South African Journal of Science

volume 114
number 11/12

Development pathways for reducing inequality and carbon emissions

Draft 2018 White Paper on Science, Technology and Innovation

Recent emergence of CAT5 tropical cyclones in the South Indian Ocean

Potential of marula waste for the production of vinegar

Econometric models to understand unemployment in South Africa

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

Volume 114
Number 11/12
November/December 2018

South African Journal of Science

eISSN: 1996-7489

Published by
the Academy of Science of South Africa (www.assaf.org.za) with financial assistance from the Department of Science & Technology.

Design and layout
SUN MeDIA Bloemfontein
T: 051 444 2552
E: publish@sunbloem.co.za

Correspondence and enquiries
sajs@assaf.org.za

Submissions
Submissions should be made at www.sajs.co.za

Review Article

Research Article

Research Letter

Cover caption
Tropical cyclone Dineo – a CAT1 storm – just off the coast of Mozambique on 15 February 2017 at 6:15 a.m. EST (1115 UTC), captured by NASA's Aqua satellite (image credit: NASA Goddard MODIS Rapid Response Team). In an article on page 83, Fitchett explores the recent emergence of CAT5 tropical cyclones in the South Indian Ocean.

Rapid change, no simple solutions

In his most recent book (*21 Lessons for the 21st Century*, Jonathan Cape, 2018), Yuval Noah Harari sets out what he believes to be, as the title suggests, the 21 most critical lessons for this century, in the context of the current state of the world. The 21 lessons are grouped into five sections, with titles such as 'The Technological Challenge', and 'Despair and Hope'. But two issues drawn from two of the sections, have messages of significance for science and education.

In the Introduction (p. ix), Harari points out that:

> *In a world deluged by irrelevant information, clarity is power. In theory, anybody can join the debate about the future of humanity, but it is so hard to maintain a clear vision. Frequently, we don't even notice that a debate is going on, or what the key questions are.*

In what follows, I draw on the lessons titled Ignorance (15); Post-truth (17); Science Fiction (18) and Education (19) – although not in any particular order.

The main thrusts of Harari's arguments are that as a species we find it difficult to escape the 'realities' of previous eras, especially those of the Industrial Revolution and its various mutations and physical structures; and that we have a tendency to replace them with false representations of what a future world (or worlds) might be like – the worlds, as he puts it, that are presented in science fiction movies.

On one hand, we struggle to move into the rapidly changing world that is happening around us, while on the other we are given fictions about the future that is emerging. Clearly these are generalisations, but they are prevalent and persistent enough to influence everything from the pattern of emerging populist politics that foregrounds nationalism, to our ignorance of the kind of world that scientific revolutions are capable of offering.

Two recurring themes in his book are (1) the benefits and other implications of scientific and technological changes of a radical nature, barely imaginable 10 years ago; and (2) archaic and increasingly irrelevant approaches to education and the related blizzard of information (and fake information) now easily accessible to anyone with a smartphone.

Starting with schooling, he points out that, for the greater part, schools are physically and intellectually still modelled on the lessons and practices of the Industrial Revolution era (yet we are already on the brink of the fifth), particularly, the production line. So, many schools still look like marginally benign factories in which children learn along a production line of facts – where millions of 'pieces', right or wrong, already abound. Two of the most important reasons why this form of schooling is not just inappropriate but wrong, lie in the fact that the world is awash with facts, and that change is now happening at a pace of rapidity previously unknown and – because of the rapidity and often unexpected nature of scientific development – largely unpredictable.

What is needed, then, apart from the essential 'three R's' is not more facts or accessible information, but the 'four C's': critical thinking, communication, collaboration and creativity. These four C's are essential in facing the reality that strangeness and the unexpected are already the new normal, as is the fact that rapid change is our only constant. But teaching both the R's and the C's is not easy (least of all in production-line schools) and so must continue, at ever more demanding levels, in higher education. For without them, the three R's become irrelevant in a world in which seeking development, discerning between fact and post-fact, and making the most of change, are core needs.

However, it is not just the school system that militates against the needed changes. Politics and politicians play their own role, as is increasingly being seen in more and more countries. The reason why these movements are problematic is that they rely on the growth of, and support for, groupthink, mostly based of fake information, but powerful enough to lead to the situations currently prevalent in England, the USA, Brazil, Turkey, Russia and Hungary, amongst others. Groupthink not only flourishes on the basis of post- or fake-fact, but also militates formidably against questioning, freedom of thought, and the four C's practised by individuals.

And yet science and technology survive and thrive, even as political movements control school learning and undermine freedom of thought and intellectual creativity, ban university courses and arrest or dismiss academics (5800 in Turkey alone), or deny hard scientific realities such as global warming.

The war against mosquitoes and malaria is steadily being won; sophisticated algorithms have changed the world of finance in radical ways; we know more about our human origins, and more accurately, than ever before; and the algorithms that drive artificial intelligence (AI) are changing the ways in which a host of human activities are undertaken – mostly successfully. But even in these positive areas, the absence of the four C's represents other major challenges. Revolutions in biotechnology and information technology made by engineers and others emerge in the context of the designers frequently knowing very little about the ecological and political implications of their successes. Without individual freedom to think and to apply the four C's, what happens to the ability to encounter and assess ethical issues such as the differences between right and wrong, or what is just and what unjust?

On balance, while change is rapid and unpredictable, while science and technology are changing the world and the ways in which we do things and live our lives, two factors are creating major problems: an education system that is, in large part, not keeping pace with the needs of young (and not so young) people; and political expediency which prefers populations whose behaviours are post-truth based and who fear individual, critical and considered thinking.

There are no simple solutions, of course. But if any one need emerges, then it is for enlightened scientists to engage more fully and consistently in debates in the public spheres of politics and education.

HOW TO CITE:
Butler-Adam J. Rapid change, no simple solutions. S Afr J Sci. 2018;114(11/12), Art. #a0299, 1 page. http://dx.doi.org/10.17159/sajs.2018/a0299

Draft White Paper on Science, Technology and Innovation neglects to prioritise issues of performance and human capability

AUTHORS:
David Walwyn[1]
Laurens Cloete[1]

AFFILIATION:
[1]Department of Engineering and Technology Management, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
David Walwyn

EMAIL:
david.walwyn@up.ac.za

KEYWORDS:
national system of innovation; knowledge production; transformation; technology transfer

HOW TO CITE:
Walwyn D, Cloete L. Draft White Paper on Science, Technology and Innovation neglects to prioritise issues of performance and human capability. S Afr J Sci. 2018;114(11/12), Art. #5679, 6 pages. https://dx.doi.org/10.17159/sajs.2018/5679

PUBLISHED:
27 November 2018

The release for public comment of the Draft 2018 White Paper on Science, Technology and Innovation[1] marks the culmination of a lengthy internal process within the Department of Science and Technology (DST). As noted in the Minister's Foreword to the Draft White Paper, the document is intended to replace both the 1996 White Paper on Science and Technology[2] and the 2008 Ten-Year Innovation Plan[3]. Its publication is the outcome of a strategic project initiated and driven several years ago by the then Minister of Science and Technology, Naledi Pandor, which has involved several cycles of input from members of the DST and its associated entities, such as the National Intellectual Property Management Office and the National Advisory Council on Innovation, and wide consultation with external experts and consultants.

Inevitably, although the draft document is relevant and broad-ranging, it carries the scars of this consultative process. There are inconsistencies (e.g. the use of terms for human development), there is a lack of detail in certain key areas (e.g. public sector innovation and new funding sources) and there is an impossibly long list of policy interventions (26 policy intents and 340 policy actions/sub-actions). (These numbers were obtained by coding the relevant chapters with ATLAS.ti and then manually filtering the coded quotes to remove statements about the future without reference to a specific action.)

Of greater concern, however, is that the document fails to ignite a convincing sentiment that science and technology can indeed contribute to the solution of South Africa's social and economic challenges. In our opinion, the following are needed to strengthen the weaknesses, before finalisation of the document:

- a clearer articulation of, and strategy for, the development of human capability and its link to economic development;
- a more critical perspective on the institutional reform, particularly public research institutions, as a means of raising the productivity of knowledge production;
- a greater emphasis on policy experimentation as a channel of transformative change, the latter in the interests of inclusivity and sustainability;
- a definitive statement on how funding will be increased and to what extent; and finally
- a much more direct list of interventions linking science and technology to economic growth and employment (i.e. a clearer and more logical theory of change), which highlights the importance of technology transfer.

More details on each aspect are given after the general overview.

General overview, core objectives and policy shifts

The core theme of the document is the accelerated deployment of science, technology and innovation (STI) in the pursuit of greater inclusivity, transformation and development, captured by the vision of 'science, technology and innovation enabling sustainable and inclusive development in a changing world'[1]. In order to achieve this vision, the DST, it is proposed, will adopt a general approach of expanding what has worked, proposing new approaches where necessary, taking advantage of opportunities presented by megatrends and promoting inclusivity and transformation.

In broad terms, these statements are irrefutable as strategies, but empty on important detail. The latter is partly contained in the more specific sections and policy proposals covering, for instance, how public institutions will be transformed or how STI will be used to accelerate economic growth. In our analysis of whether the proposed instruments adequately address the objectives, we have constructed a classical policy matrix, as shown in Table 1.

The matrix allows for a more detailed critique of objectives versus instruments. The latter are categorised into the three-fold typology of regulatory instruments, financial and economic instruments, and soft instruments, which are referred to as the 'sticks, carrots and sermons' of public policy.[4] This typology, which allows for the grouping of policy instruments into a limited number of well-defined categories, has been effectively applied to innovation policy as a means of understanding and designing suitable policy mixes.[5]

It is apparent from the matrix that the objective of 'policy coherence and coordination' overwhelms all other priorities, with implementation relying on the soft instrument of intra-government coordination, consultation and planning processes. The more critical objectives of enhancing economic growth, developing human capability and improving funding are mentioned less frequently, and sustainability is almost completely neglected. Table 1 and its analysis support our listing of the main policy gaps.

Table 1: The policy matrix of the Draft White Paper, constructed using the conventional categories for policy instruments[5]

Policy objective	Policy instrument										
	Regulation			Economic transfer				Soft instruments			
	Review intellectual property rights	Reform institutions	Adapt other legislation	Change funding allocation	Use public procurement	Invest in human capability	Incentivise other investment	Encourage collaboration	Plan and consult	Effect intra-governmental coordination	Monitor and evaluate
Human capability development			1	5		18		5		8	2
Greater inclusion and transformation	3			9		8	2	4	3	5	1
Sustainability			1	5					1	1	
Enhanced economic growth		3	6	23	4	2	2	4	5	6	
Improved partnerships (NSI)		2		10	2	1	2	19	11	12	
Policy coherence and coordination	1	3	3	6	3	3	1	7	29	38	14
Performance (of public institutions)		3		3			2	4		5	5
Expanded NSI and research enterprise		2		19	1	1	1	3	5	5	1
Enabling innovation environment	4		6	15	2		1	1	6	6	1
Public sector innovation								1		2	
Improved NSI funding regime			1	15	5		4	2	13	10	2

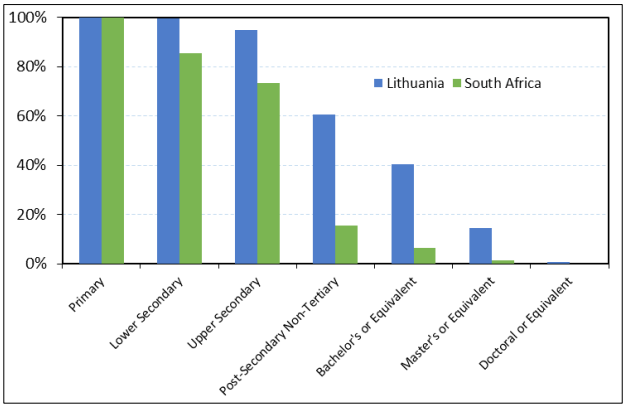
Note: *The table lists the number of times that a particular policy objective is mentioned in the form of a specific policy action. In the construction of the matrix, we considered both the explicit objectives in the Draft White Paper as well as two implied objectives, namely sustainability and public sector innovation. We have omitted policy instruments that occur fewer than eight times, as well as cases in which the nature of the instrument is not clear or in which there is no obvious link to a policy objective. Some planned policy actions support multiple objectives, and some implement more than one policy instrument. The 99 sub-actions identified were not included in this analysis; only the 340 policy actions were tabulated.*

Policy gaps

Human capability

Despite being acknowledged over a long period as being a core weakness, the necessary development of human capability in South Africa remains unrealised.[6] Although the Draft White Paper has a chapter on human capital, its placement in the document and its title should be changed, reflecting not content editing but a performative and semiotic imperative. Human capability is fundamental – not only to economic growth, but also for the materialisation, at the level of the individual, of ethical values, citizenship, and specific goals of pursuit which align with principles of social and environmental justice.[7]

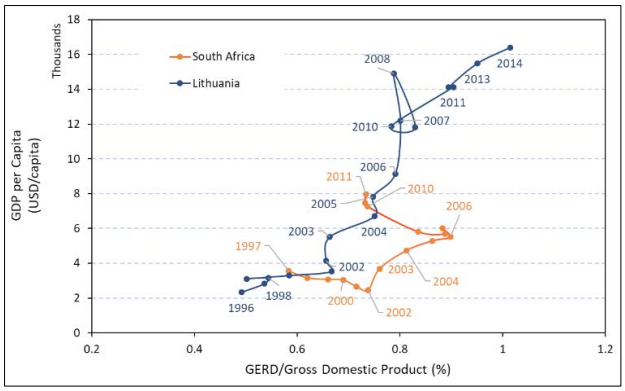
The benefits of human development are well illustrated by a comparative study of post-Soviet Baltic, Central Asian and Transcaucasian states. Lithuania in particular is a post-Soviet success story, driven by a high level of human capability which remained in place after the collapse of the Soviet Union. Already in 1994, the country had a literacy rate of 99%, but by 2016 more than 60% of the population had a post-secondary qualification, as shown in Figure 1.



Source: OECD[30]

Figure 1: Education attainment levels in Lithuania and South Africa, 2016.

In combination with higher levels of gross expenditure on research and development, this emphasis on human capability development has had a radical impact on economic development and per capita income, as shown in Figure 2.



Source: World Bank[31]

Figure 2: Gross expenditure on research and development (GERD) and per capita income for South Africa and Lithuania, 1994 to 2016.

The Draft White Paper employs a confusing medley of terms relating to human development. Our view is that all references to human resource development (used 19 times) and human capital (used 8 times) should be replaced; these terms are narrow and neglect the real importance of human development, and particularly post-secondary education. As a policy document, the White Paper needs to set a new precedent in terms of how education is conceptualised, and hence how the roles of public educational institutions are defined. In particular, the development of human capability at post-secondary level should be identified as the first and most urgent priority.

Management of higher education and public research institutions

The development of the Draft White Paper has been preceded by an impressive number of studies and reports[8-13], most of which have recommended the transformation of public sector innovation-linked institutions. The Draft White Paper has responded to these recommendations through a number of broad policy proposals, including the establishment of an Inter-Ministerial Committee on Science and Technology, revision to the mandate of the National Advisory Council on Innovation and strengthening the governance of public research institutions.

In our view, these proposals are vague and non-committal. The intentions of reform are limited to reducing overlap and inefficiencies, and even expanding the number of institutions. A recent review of public-funded research and development (R&D) highlighted the poor performance of the science councils and the intramural government research institutions relative to the universities as producers of many forms of research and innovation outputs, including scientific publications, patents, spin-off companies, contract research income and research qualifications.[13] The review adopted an approach which monetised the various types of outputs and then calculated an overall return on investment from the public funding, the results of which are shown in Figure 3. It was concluded that the universities, as presently configured and assessed, represent the most attractive return for public funding.[14]

Concerns have already been raised about the high cost and low output, relative to their mandate, of the science councils[12], suggesting a much more radical approach to the restructuring of public research institutions, including the following:

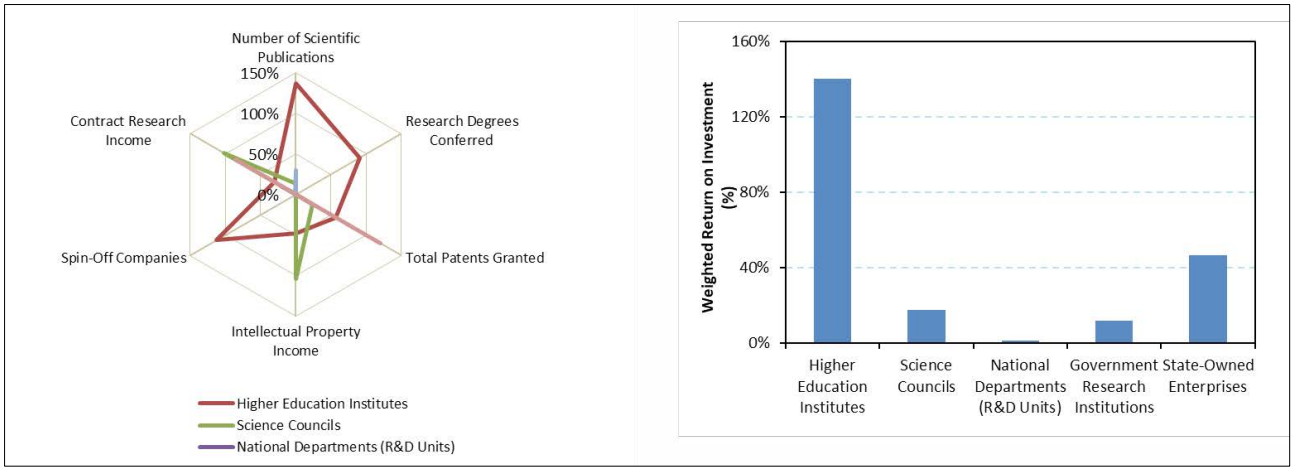
- closure of the Human Sciences Research Council, with its core units being moved to the universities or, in the case of the Centre for Science, Technology and Innovation Indicators, to the National Advisory Council on Innovation;
- separation of the National Facilities (essentially performance agencies) from the National Research Foundation (a funding agency), and the general integration of such facilities where long-term capital investment is required; and
- separation of the funding and performance arms of the Medical Research Council, with the funding portfolio being moved to the National Research Foundation.

We agree with the Draft White Paper intent of the development of a policy framework to 'describe the purpose, functions and governance of public research institutions', and feel that this aim should be undertaken as a matter of urgency. A critical component of this framework should be the clear definition of the rationale for public research institutions, to allow for the more efficient allocation of government's research needs, in cases in which capital-intensive research infrastructure is not required, from science councils to the universities.

Policy experimentation and policy mix

Policy mix and policy experimentation have emerged since the publication of the 1994 White Paper on Science and Technology as important developments in the field of innovation studies. Experimentation in innovation policy recognises that the national system of innovation (NSI) is a complex system, and that policy outcomes are often uncertain, especially in developing countries.[15] Experiments permit the exploration of new approaches, particularly in addressing wicked or intractable problems, in a dynamic and positive style, supporting the principles of reflexive governance and enabling the establishment of niches which can be scaled to broader programmes more reliably.[16-18]

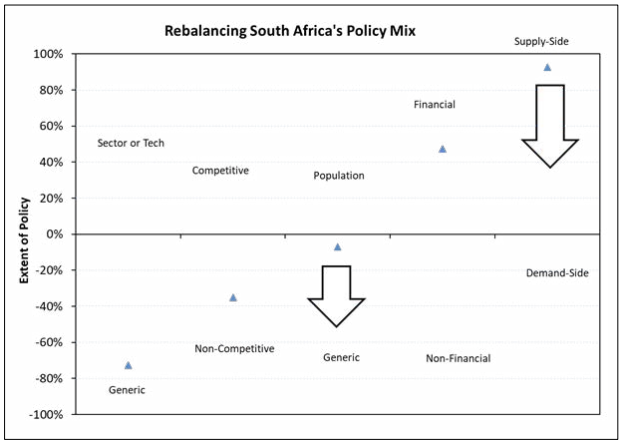
Similarly, the rationale for considering policy mix, rather than individual policy instruments, is several-fold; in the first instance, policy beneficiaries are diverse and require different approaches to achieve



Source: RebelGroup South Africa[13]

Figure 3: Comparison of the outputs from public-funded R&D at public institutions.

similar outcomes.[19] Furthermore, policies themselves interact and are interdependent, requiring a more holistic approach to policy design and analysis.[20] It is argued that a policy mix approach is particularly important in addressing the objectives of socio-technical transformation and sustainability transitions[21], both of which are core intents of the Draft White Paper. In our own research, we have shown that South Africa's innovation policy mix is dominated by supply-side measures, supporting early stage R&D but providing more limited assistance for market development.[19] Rebalancing the innovation policy mix towards the use of more demand-side instruments (such as the use of public procurement as a means of stimulating innovation), combined with generic rather than population targeted policies, as shown in Figure 4, could improve policy outcomes.



Source: Naidoo[19]

Figure 4: Recommended adjustments to South Africa's innovation policy mix.

It is our view that the DST is missing an opportunity to mainstream policy experimentation and mixes in its Draft White Paper. A specific section in the document covering the importance of both approaches as a means of addressing the considerable and intractable system problems, such as innovation-led economic growth, would be invaluable in both introducing these methods as legitimate processes within government, and also to improve upon the document's underlying theory of change.

Indeed, in its present form, there is no explicit theory of change. Although some of the specific policy intents are linked to the desired outcome (e.g. the establishment of the Inter-Ministerial Committee on Science and Technology as a means of improving policy coherence), the overall theory is not stated, nor is it apparent in the overview material. Policy experimentation itself assumes a particular theory of change, namely that such experiments lead to broader systemic change as a consequence of scale-up from the niche (micro) level to new regimes and eventually new landscapes.[17] It is also more amenable to implementation, which, as we know, requires the synchronicity of an acknowledgement of the problem, the existence of an appropriate policy and acceptance by politicians of the proposed policy solution.[20,22] Experimentation allows implementation to proceed even if there is still some doubt about the immediate prospects of a solution.

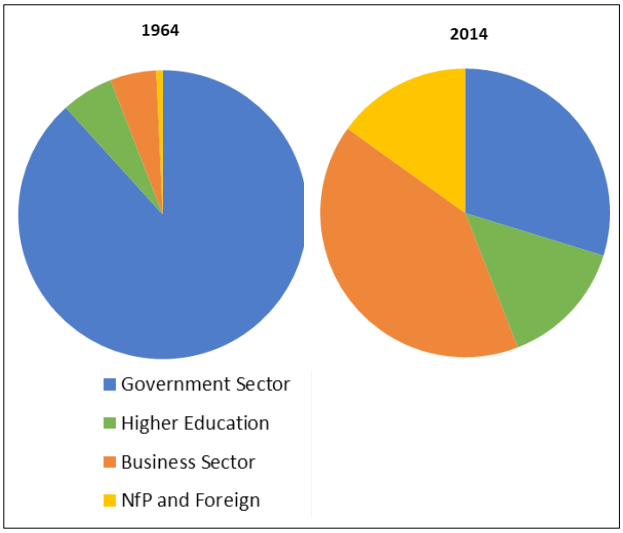
Economic growth led by technological change

The award of the 2018 Nobel Prize for Economics to Paul Romer, in acknowledgement of his work on endogenous growth theory, further strengthens claims of many economists and innovation theorists that economic growth is directly linked to technological innovation.[23] Romer[24] noted that the important sources of economic growth are technological change and increases in 'human capital', which are in turn driven by intentional investment decisions in R&D and other sources of innovation.

This insight was derived from data in developed countries and should be mediated by the South African context in two respects. Firstly, the most important path for technical change in developing countries is technology transfer and diffusion, broadly described as innovation by 'doing, using and interacting (DUI)'[23]. Secondly, the two important determinants of successful DUI are a strong absorptive capacity within the business sector and public research institutions, and high-level policy support for technology transfer.

Absorptive capacity depends on a complex set of antecedents including the two important supply-side factors of human capability and public-funded R&D, and the three demand-side elements of legitimacy, local market and entrepreneurial activity.[25,26] The Draft White Paper implicitly acknowledges the important role of public-funded R&D in stimulating economic growth, and the now widely accepted perspective that such R&D has a higher level of return than private R&D[27], so long as knowledge systems are sufficiently open to ensure knowledge diffusion and economic spillovers.

As a result, the Draft White Paper highlights the need to expand present levels of funding but fails to indicate how this expansion can be achieved under the present economic conditions of contraction and fiscal restraint. Targets for gross expenditure on R&D are frequently not met and have been described as 'wishful thinking'[28]. South Africa is no exception, having consistently failed to meet its own intensity goals.[10] In a climate of many competing priorities, it is difficult to imagine how additional funding will be secured, but it is our opinion that government needs to lead the trend through reprioritisation of existing budget allocations. There is precedence, even in South Africa, for a more expansive role of the state in R&D; a longitudinal study of historical expenditure has shown that in 1964, public funding of R&D accounted for 80% of the total expenditure (Figure 5). Although the context is very different, it could be argued that government R&D funding in 1964 enabled the country's subsequent high rate of economic growth.



Source: Van Wyk et al.[32]

Figure 5: Profile of R&D expenditure in South Africa by source of funds, 1964 and 2014.

In terms of high-level support, the Draft White Paper does seek to include such a goal, and hopefully it will find a stronger reception than in previous administrations. In this area, South Africa should learn from China's science and technology policy, which has over a long period advocated and implemented a highly proactive role for the state in technology transfer and R&D as the starting point for its innovation-driven development and economic growth strategy. Indeed, China's approach to science and technology has been unequivocal and completely unambiguous. For instance, on many occasions, President Xi Jinping and his predecessors have stressed the importance of innovation in economic growth, declaring that 'innovation is the most powerful lever for development' and the need to 'strive for both scientific and technological innovation, and institutional innovation … to fully unlock our development potential'[29].

Discussion and conclusion

The challenges of policy processes are considerable. Policy should be consulted, not just in the interests of democracy and participation, but also to ensure alignment, policy coherence and stakeholder support, and to avoid any disastrous, unintended consequences. However, consultation may lead to a dilution of focus, a lack of clear prioritisation and the rallying of resistant elements which can impede system-wide necessary change. Important aspects of effective policies – such as being able to reallocate resources and to effect the necessary transformations – can be annulled by consultative processes.

We argue that rationalisation of the policy actions through a clearer theory of change is essential. In this respect, the NSI approach may have outlived its usefulness. Its adoption as a guiding framework for innovation policy in South Africa was a political, not a technocratic, perspective. The approach relied upon an ambitious level of agency at the micro (or individual) level, and a high level of efficiency at the meso (firm and government department) level. Agency depends on human capability which, as reported in many studies, is an area in which South Africa generally fails dismally. Moreover, meso-level performance is patchy in both public and private sectors, with the last decade of state patrimonialism being disastrous for South Africa's economy and the NSI.

In the absence of widespread agency and efficiency, the NSI framework may not be a sufficiently radical approach to achieving the broader goals of the Draft White Paper. In our view, although the application of the NSI framework has, so far, been insightful and constructive, and there is now a broader consensus within government about innovation-led growth, the framework's political assumptions are too conservative about the role of its actors, and could be changed to the more experimental but transition-based approach of the multi-level perspective.[17]

Our core advice to the policy architects is to abandon the more generic platitudes of NSI theory and to strengthen the institutions (used in the sense of laws, regulations and codes) and organisations of the state in their role as agents for innovation and technology transfer. Make human capability development the top priority, close non-performing science councils, ensure the clear separation of funding and performance mandates, and institutionalise policy experiments as a means of achieving transformation, inclusivity and sustainable development. In this way, the comments of the Deputy Minister in her Foreword to the Draft White Paper may indeed be prescient:

> *I am confident that through efficient implementation of this new STI policy by various stakeholders in the public and private sector, the lives of our people and the fortunes of our communities will be transformed through STI.*

References

1. South African Department of Science and Technology (DST). Draft White Paper on Science, Technology and Innovation. Pretoria: DST; 2018 [cited 2018 Sep 26]. Available from: https://www.dst.gov.za/images/2018/Draft-White-paper--on-STI-7_09.pdf
2. South African Department of Arts, Culture, Science and Technology (DACST). White Paper on Science and Technology. Pretoria: DACST; 1996.
3. South African Department of Science and Technology (DST). Ten year innovation plan: Innovation towards a knowledge-based economy 2008–2018. Pretoria: DST; 2008.
4. Bemelmans-Videc M-L, Rist RC, Vedung E. Carrots, sticks and sermons: Policy instruments and their evaluation. London: Transaction; 2003.
5. Borrás S, Edquist C. The choice of innovation policy instruments. Technol Forecast Soc Change. 2013;80(8):1513–1522. https://doi.org/10.1016/j.techfore.2013.03.002
6. United Nations Development Programme. Human development indices and indicators: 2018 Statistical update, South Africa. Geneva: United Nations Development Programme; 2018.
7. Walker M, Fongwa S. Universities, employability and human development. London: Palgrave Macmillan; 2017. https://doi.org/10.1057/978-1-137-58452-6
8. Walwyn D. Synthesis report: Review of the White Paper on Science and Technology and high level framing for a new decadal plan. Pretoria: National Advisory Council on Innovation; 2016.
9. South African national survey of intellectual property and technology transfer at publicly funded research institutions. Pretoria: DST, SARIMA, NIPMO, CeSTII; 2017. Available from: http://www.hsrc.ac.za/en/research-outputs/view/8578
10. Ministerial Review Panel. Ministerial Review Committee on science, technology and innovation landscape in South Africa: Final report. Pretoria: Department of Science and Technology; 2012.
11. Ministerial Review Panel. Ke Nako: A review of the South African science, technology and innovation institutional landscape. Pretoria: Department of Science and Technology; 2017.
12. Bertoldi A, Gardner D, Hague K, Lockwood K, McGloughlin R, Walwyn D. Assessment of the effective partnering of science councils with the private sector: Findings and recommendations. Pretoria: National Treasury; 2014.
13. RebelGroup South Africa. Review of the funding systems, mechanisms, and instruments adopted by government in funding for research, development and innovation including an international bench-mark analysis. Pretoria: National Treasury; 2018.
14. Walwyn D, Cloete L. Universities are becoming major players in the national system of innovation. S Afr J Sci. 2016;112(7–8), Art. #2015-0358, 8 pages. http://dx.doi.org/10.17159/sajs.2016/20150358
15. Husain L. Policy experimentation and innovation as a response to complexity in China's management of health reforms. Globalization Health. 2017;13(1), Art. #54, 13 pages. https://doi.org/10.1186/s12992-017-0277-x
16. Kivimaa P, Hildén M, Huitema D, Jordan A, Newig J. Experiments in climate governance – A systematic review of research on energy and built environment transitions. J Clean Prod. 2017;169:17–29. https://doi.org/10.1016/j.jclepro.2017.01.027
17. Schot J, Geels FW. Strategic niche management and sustainable innovation journeys: Theory, findings, research agenda, and policy. Tech Anal Strat Manag. 2008;20(5):537–554. https://doi.org/10.1080/09537320802292651
18. Hildén M, Jordan A, Huitema D. Special issue on experimentation for climate change solutions editorial: The search for climate change and sustainability solutions – The promise and the pitfalls of experimentation. J Clean Prod. 2017;169:1–7. https://doi.org/10.1016/j.jclepro.2017.09.019
19. Naidoo S. Rebalancing innovation policy mix to improve support for South Africa's manufacturing sector [MBA mini dissertation]. Johannesburg: Gordon Institute of Business Science, University of Pretoria; 2017.
20. Flanagan K, Uyarra E, Laranja M. Reconceptualising the 'policy mix' for innovation. Res Policy. 2011;40(5):702–713. https://doi.org/10.1016/j.respol.2011.02.005
21. Kivimaa P, Kern F. Creative destruction or mere niche support? Innovation policy mixes for sustainability transitions. Res Policy. 2016;45(1):205–217. https://doi.org/10.1016/j.respol.2015.09.008
22. Kingdon JW, Thurber JA. Agendas, alternatives, and public policies. Boston, MA: Little and Brown; 1984.
23. Gries T, Grundmann R, Palnau I, Redlin M. Innovations, growth and participation in advanced economies – A review of major concepts and findings. Int Econ Econ Policy. 2017:1–59. https://doi.org/10.1007/s10368-016-0371-1
24. Romer PM. Endogenous technological change. J Polit Econ. 1990;98(5):71–102. https://doi.org/10.1086/261725
25. Walz R. Towards a dynamic understanding of innovation systems: An integrated TIS-MLP approach for wind turbines. In: Horbach J, Reif C, editors. New developments in eco-innovation research. Cham: Springer; 2018. p. 277–295. https://doi.org/10.1007/978-3-319-93019-0_13
26. Kebede KY, Mitsufuji T. Technological innovation system building for diffusion of renewable energy technology: A case of solar PV systems in Ethiopia. Technol Forecast Social Change. 2017;114:242–253. https://doi.org/10.1016/j.techfore.2016.08.018
27. Griliches Z. R&D and productivity: Economic results and measurement issues. In: Stoneman P, editor. Handbook of the economics of innovation and technological change. Cambridge, MA: Blackwell; 1995. p. 52–89.
28. Carvalho A. Wishful thinking about R&D policy targets: What governments promise and what they actually deliver. Sci Public Pol. 2018;45(3):373–391. https://doi.org/10.1093/scipol/scx069

29. Miesing P, Tang M. Technology transfer institutions in China: A comparison of value chain and organizational structure perspectives. In: Siegel DS, editor. The world scientific reference on innovation 1. Singapore: World Scientific; 2018. p. 43–60. https://doi.org/10.1142/9789813149045_0003

30. Organisation for Economic Cooperation and Development (OECD). Education data 2016. Paris: OECD; 2018 [cited 2018 Oct 08]. Available from: https://data.oecd.org/education.htm.

31. World Bank. World development indicators (2018). Washington: World Bank; 2018 [cited 2018 Mar 15]. Available from: http://databank.worldbank.org/data/home.aspx

32. Van Wyk RJ, Mawby EL, Eaton WL. Expenditure on research and development in the natural sciences undertaken within the government sector and by universities in South Africa during the financial years 1964/65 and 1965/66, and the academic years 1964 and 1965. Pretoria: Industrial Economics Division, CSIR; 1968.

On the power of restraint in the writing of lives: Humanities Book Award 2018

AUTHOR:
Jonathan D. Jansen[1]

AFFILIATION:
[1]Distinguished Professor, Faculty of Education, Stellenbosch University, Stellenbosch, South Africa

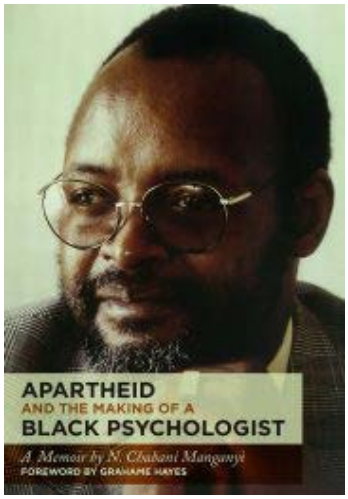
CORRESPONDENCE TO:
Jonathan Jansen

EMAIL:
jonathanjansen@sun.ac.za

KEYWORDS:
psychology; autobiography; ideas; apartheid; South Africa

HOW TO CITE:
Jansen JD. On the power of restraint in the writing of lives: Humanities Book Award 2018. S Afr J Sci. 2018;114(11/12), Art. #a0295, 2 pages. https://dx.doi.org/10.17159/sajs.2018/a0295

PUBLISHED:
27 November 2018



It is a wisdom attributed to the Greek historian Thucydides, that 'Of all manifestations of power, restraint impresses men most'. To read Chabani Manganyi's *Apartheid and the Making of a Black Psychologist* is to be awed by the power of restraint in the hands of a capable author and a compelling thinker. Restraint, in this context, is not synonymous with 'not telling' or 'holding back' on writing about a troublesome event. In the hands of a highly skilled writer, restraint is a way of communicating where the very art of control over the writing enables the telling to explode off the pages of the book.

Since the end of apartheid there has been a veritable industry of autobiographies reminding readers exactly how bad the past was or how its terrible legacy hovers in the present. Everyone has a story to tell, from exiles to 'inziles', from military officers of the white regime to soldiers of the liberation movements, from heroes of the struggle to memoirs of the 'born free'. In all these books the prose is unsubtle, sometimes crude, and often angry. This kind of biographical writing is so common that Manganyi's book surprises with its insight, subtlety and restraint.

What made this scholarly masterpiece, published by Wits University Press in 2016, the unanimous choice for the 2018 ASSAf Humanities Book Award – among 38 competitors – is its uniqueness in another respect. This contribution is, in essence, a history of ideas told through the life story of a black psychologist since the 1940s. Skilfully, the account of Manganyi's life is intertwined with his evolving ideas about the human condition and the growth of ideas in his specialist field of psychology – all of which is done against the backdrop of the advent and unfolding of apartheid from 1948 onwards.

It is painful reading the vivid account of the impact of the migrant labour system on his father and the family. Without rancour, Manganyi shows exactly how migrant work sought to sustain fragile domestic economies even as it separated families and stripped self-respecting African men of dignity, reducing them to 'an unmistakeable sense of shame, silent anger and regret' (p. 11).

Manganyi goes on to account for his life in a mission boarding school called Lemana in today's Limpopo Province and his passage through university as a student and later as a professor and head of psychology at the Umtata Campus of the University of Fort Hare. In briefly referencing the story of these two institutions – Lemana and Umtata – Manganyi offers a first-hand account of how apartheid transformed powerful black institutions into intellectual wastelands. He remembers how an 'intellectual bounty of excellent teachers' (p. 12) at school would be cancelled out by Bantu Education and how a campus with lecturers of 'outstanding academic promise' (p. 56) would be flattened by the new homeland system that brought into existence the University of Transkei. Manganyi is correct that these transformations of institutions by apartheid in the 1950s have yet to be more fully accounted for in scholarly histories of education, especially in the case of Fort Hare's Umtata Campus that stood on the brink of building from the ground up a place of academic excellence.

It is, however, in his search for a job that the systematic racism of organisations becomes evident – from his short stay as 'Employee Relations Officer' at Ellerines furniture company to the denial of his request for internship training in clinical psychology settings in the only place that offered such opportunity, white hospitals. Tongue-in-cheek, Manganyi quips that 'the most difficult hurdle, and one that could not be easily overcome, was the colour of my skin' (p. 26). So he ends up at Baragwanath Hospital – as a clinical psychologist in training in a neurosurgery department – and is housed in the black doctors' quarters. Time after time his career was sabotaged by the race laws of apartheid and the bigotry of businesses (Ellerines pushed him out because he pointed to racism as the reason for the high turnover of black staff) and institutions such as Wits which could not get itself to appoint this brilliant black scholar as anything more than a 'Visiting Professor' – a shame the Johannesburg university has never corrected.

This is where the foundation of Manganyi's intellectual orientation is revealed. He is no victim of the apartheid's suffocating 'laws and practices that followed me relentlessly' (p. xiv). Manganyi is equipped with an unusual curiosity about life, such as that expressed in his seminal publication *Being-black-in-the-world*. Time after time he turns frustration into a quest for deep understanding. A vicious caning for absenteeism from school is a chance for reflection: 'I still wonder what would have happened if the principal had spared the cane' (p. 3). The sub-standard black doctors' quarters at Baragwanath offered insight into 'the social and intellectual culture of a racially segregated teaching hospital' (p. 27). Manganyi was haunted by the horrific necklacing of the suspected police informer, Maki Skosana, which launched him into a funded study on political violence. And as an expert witness defending activists from the death penalty, he was intrigued by the biographies of those accused of political murders and brought these richly drawn contexts to the attention of the courts for mitigation of sentence.

At this point, a personal but related note. It was Professor Manganyi who advised the then Rector of the University of Pretoria to recruit me as Dean to this formidable Afrikaans institution. Each Dean reported to a Vice-Rector

and I was lucky to at one stage report to Manganyi. Rather than talk academic business and administration, our weekly sessions turned out to be two-man seminars on complex questions of campus and country. One day I came for my usual consultation and complained bitterly about race, hierarchy and bureaucracy at the university. My mentor listened for a while and then, in his typically calm and reserved manner, said something that changed the direction of my own scholarship. 'JJ' he mused, 'your problem is that you get angry before you think. Try and make sense of what is going on in this place.' That was the moment in which the idea for my book, *Knowledge in the Blood: Transforming Race and the Apartheid Past*, was born. He had taught me an invaluable lesson: try to understand deeply that which frustrates you, for then, and then only, can you change it.

A demeanour of restraint did not mean that Manganyi was not at various point(s) angry, frustrated and even embittered by the traumatic experiences of his life. But he expressed it through his scholarship, as in the semi-fictional work *Mashangu's Reverie*, a work of self-analysis which 'started to rid myself of disturbing impulses' (p. 47). It was when he was out of the country, on a scholarship to Yale University, that Manganyi found himself free to be angry, to unburden himself through writing and to contemplate the possibility of life in exile – a contemplation which would complete a trilogy of biographies on the lives of exiled South Africans.

The golden thread binding together the various moments in Manganyi's compelling life story is ideas. The mission school gave access to 'the world of knowledge' (p. 14) and 'discoveries of science' (p. 13) despite its colonial bent. The University of the North offered access to the disciplines by knowledgeable (white) lecturers and an assistantship opportunity in educational psychology despite the ideological distortions of apartheid education. The honours degree in psychology at Unisa provided a 'most stimulating' (p. 19) exposure to studies in the philosophy of science. Then followed a master's degree drawing on Maslow's theory of human needs among factory workers and furniture salesmen using statistical methods; there was much frustration with the unresponsive white supervisor but 'I did not twiddle my thumbs in confusion and self-pity' (p. 24), recalls Manganyi who soon qualified for doctoral studies.

Despite these solid foundations in clinical psychology, discrimination meant pursuing ideas on his own and through opportunities in non-ideal places such as the neurology department at Baragwanath and without specialist supervision in psychiatry. Even as an expert witness in apartheid courts, he was practising forensic psychology when the field was in its infancy. 'But' says Manganyi, 'I kept my eyes and ears open' (p. 29) and 'I read myself into important but unfamiliar knowledge domains' (p. 30). In other words, his formal academic and clinical training came through a combination of structured teaching and professional neglect out of which he crafted, through reading and observation, the knowledge necessary for success as a psychologist – such as his forays into the study of hysteria. It was the opportunity to study at Yale as a postdoctoral fellow that enabled Manganyi to escape 'the minimalist hit-and-miss experience I had to create for myself at Baragwanath' (p. 40). Here he would gain thorough knowledge in a range of clinical psychology skills such as psychoanalytic psychotherapy. Returning home, Manganyi's ideas returned him to a life-long interest in biography, capturing the lives of Eskia Mphalele, Dumile Feni and Gerard Sekoto in powerful accounts of the exiled condition.

Strikingly, Manganyi has been able to maintain an active research and publication schedule throughout his career in academic leadership in psychology, governmental leadership as Director General for Education, university leadership as Vice Chancellor and executive leadership of a major private sector funded organisation. A man of ideas, he never abandons his intellectual pursuits, always asking hard questions about the human condition.

Chabani Manganyi is among the last of a special class of South African intellectuals who treasured ideas above the pursuit of wealth, status or petty politics. When told at the end of his stint as Director General for Education that he would be 'redeployed', you could inflict on this cultured man no greater insult; he declined the offer. When he engaged academic staff at the University of the North on critical matters at the university, his heart sunk as he took in the low level of academic discourse and ambition, prompting the Vice Chancellor to start planning his exit. In the world of Manganyi, ideas still mattered.

What this award-winning book accomplishes is to turn the two major intellectual preoccupations of Manganyi's professional career – psychology and biography – into a powerful work of psychobiography applied to his own life. In this sense, his work and his life have come full circle and, in the process, he teaches generations of South Africans how to be human in a broken world.

ASSAf consensus study on the ethical, legal and social implications of genetics and genomics in South Africa

AUTHORS:
Michael S. Pepper[1,2]
Collet Dandara[3]
Jantina de Vries[4]
Amaboo Dhai[5]
Melodie Labuschaigne[6]
Freddy Mnyongani[7]
Keymanthri Moodley[8]
Antonel Olckers[9]
Anne Pope[10]
Raj Ramesar[11]
Michele Ramsay[12]
Himla Soodyall[13,14]
Wayne Towers[15]

AFFILIATIONS:
[1]Institute for Cellular and Molecular Medicine, Department of Immunology, Faculty of Health Sciences, University of Pretoria, Pretoria, South Africa
[2]South African Medical Research Council Extramural Unit for Stem Cell Research and Therapy, University of Pretoria, Pretoria, South Africa
[3]Division of Human Genetics, Department of Pathology and Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
[4]Department of Medicine, University of Cape Town, Cape Town, South Africa
[5]Steve Biko Centre for Bioethics, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[6]Department of Jurisprudence, School of Law, University of South Africa, Pretoria, South Africa
[7]School of Law, University of KwaZulu-Natal, Durban, South Africa
[8]Centre for Medical Ethics and Law, Department of Medicine, Faculty of Health Sciences, Stellenbosch University, Stellenbosch, South Africa
[9]DNAbiotec (Pty) Ltd, Pretoria, South Africa
[10]Emeritus Associate Professor, Department of Private Law, Faculty of Law, University of Cape Town, Cape Town,

South Africa is home to one of the most genetically diverse populations in the world, which, combined with its high disease burden and high-quality infrastructure, makes our country a prime location for conducting genetics and genomics research. South African genomes are therefore highly sought after by the global research community. Increasingly, a range of technological advances, including the possibility of 'reading' whole genomes or exomes through next-generation sequencing, allows access to detailed molecular information from which information about health and disease can be inferred. This reading may also occur when information is collected for different purposes, which raises questions about the ethics of inferring information about health and disease in these situations. Against this backdrop of technological and ethical complexity, there is an urgent need to understand and protect the interests of patients and individuals who participate in research in the fields of genetics and genomics.

The publication of the sequence of the human genome 15 years ago heralded humankind's entry into a new era in which large volumes of data would be generated with the potential to drive major improvements in human health. A massive data management sector has evolved in parallel and includes data acquisition, storage, analysis, interpretation and access. This evolution of data has been accompanied by growth of commercial companies that utilise this information through the generation of products and services. Consequently, genetics and genomics present both opportunities and challenges in the South African context. Because of South Africa's history, during which segregation and oppression ruled – the effects of which are still experienced today – research as well as clinical and forensic practice in genetics and genomics raise many difficult issues that must be dealt with sensitively and constructively. Moreover, targeted policy, legislation, regulations or guidelines are lacking in these fields. For these reasons, in 2016, the Academy of Science of South Africa (ASSAf) undertook to conduct a consensus study on the ethical, legal and social implications (ELSI) of genetics and genomics work, as they relate to research, health service provision and forensic applications (medical and legal) in South Africa. The purpose of this study was to provide a well-researched document based on a combination of international best practices adapted to local conditions and deliberations by the panel, that will assist the national Departments of Health and Science and Technology to draft legislation, regulations and guidelines on matters pertaining to human genetics and the human genome. The study has been completed and a synopsis of some of the key issues and recommendations flowing from the consensus study is provided here.

The full report is divided into three thematic areas significant for the ELSI of genetics and genomics work: (1) building relationships, (2) respect for persons and (3) good stewardship. Following analysis of each theme, pragmatic, ethically and legally sound, culturally appropriate, feasible, enforceable and sustainable recommendations are proposed to optimise use of resources in the country. All recommendations are intended to underpin dialogue and discussions that must lead to new policy, legislation, regulations and guidelines.

Given that genetic and genomic information has implications for individuals, families and communities, the report looks at the ELSI of genetics and genomics in South Africa through the lens of the communitarian ubuntu philosophy, loosely translated as 'humanity', or more commonly 'I am because we are', which continues to drive a South African national consciousness in the process of democratic transformation. Ubuntu deepens respect for persons and pervades every sphere of South African life, including science. The degree to which ubuntu is practised consciously may differ in urban and rural communities. It may also differ amongst youth and older members of society. In this regard, it may be said that relative solidarity is an important component of ubuntu. That it is increasingly pervasive is illustrated by the observation that the South African judiciary has embraced ubuntu as an integral part of the constitutional values and principles, especially when interpreting the Bill of Rights. The panel therefore recommends that the ubuntu principle must be promoted in genetics and genomics research, healthcare delivery and forensics practice (Recommendation 5). Although the notions of personal autonomy and ubuntu may appear to be in tension, they should be seen to be complementary rather than mutually exclusive. Ubuntu does not negate individual choice or the exercise of autonomy; rather, individual choice is expected to take the community context into account. Genetics and genomics are good examples of why ubuntu should infuse our decision-making, because of the inherent 'domino-effect' of those choices on families and communities.

The section on 'Building Relationships' focuses on engagement between human genetics/genomics practitioners and the general public and communities, and recommends that close attention is given to stakeholder engagement (Recommendation 1) to promote understanding amongst all role players about their roles and responsibilities. Topics for engagement range from academic research projects, to genetic testing in the public and private sectors, and also include the relationships among the public, the criminal justice system and the forensic science sector of the country. Accountability and transparency are emphasised for all aspects (Recommendation 4). Further, the report highlights South African experiences with community engagement for genomics and the critical importance of education and translation of science into policy and practice (Recommendation 2). Regular evaluation of the effectiveness of stakeholder engagements is also recommended.

This section also emphasises the importance of ensuring that the public is well informed about participating in research projects, as well as their roles, rights and responsibilities when choosing to use direct-to-consumer genetic marketing and testing approaches (Recommendations 3 and 8). Regarding direct-to-consumer marketing, the advent of social media and the ease with which information is transmitted have facilitated access to large

South Africa

[11]MRC/UCT Research Unit for Genomic and Precision Medicine, Division of Human Genetics, Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa

[12]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[13]Division of Human Genetics, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[14]National Health Laboratory Service, Johannesburg, South Africa

[15]Africa Unit for Transdisciplinary Health Research (AUTHeR), North-West University, Potchefstroom, South Africa

CORRESPONDENCE TO:
Michael Pepper

EMAIL:
michael.pepper@up.ac.za

KEYWORDS:
regulation; stewardship; ubuntu; health care; privacy

HOW TO CITE:
Pepper MS, Dandara C, De Vries J, Dhai A, Labuschaigne M, Mnyongani F, et al. ASSAf consensus study on the ethical, legal and social implications of genetics and genomics in South Africa. S Afr J Sci. 2018;114(11/12), Art. #a0302, 3 pages. https://dx.doi.org/10.17159/sajs.2018/a0302

PUBLISHED:
27 November 2018

amounts of information on the possible consequences of genomic variants on health and disease for individuals and their families. This desire for information is driving an industry in which individuals can obtain very detailed information on their own genomes simply by providing a biological sample (e.g. sputum or a cheek swab) and mailing it to an entity that will do the testing and analysis. Although this practice promotes the exercise of autonomy to access meaningful personal information, the process could impact negatively on the person depending on (1) the quality of the test and (2) the means by which results and the implications of the findings are delivered, which should be through direct counselling. Negative effects create anxiety and confusion, as the recipient may not know how to respond to the information. Secondly, not all genetic/genomic variants have consequences for health and disease. To protect the public, the panel recommends that direct-to-consumer genetic marketing and testing must be regulated (Recommendation 3). The panel also recommends that the South African Health Products Regulatory Authority (SAHPRA) should include regulation of genetic tests under the *Medical Devices Act* (*No. 14 of 2015*) (Recommendation 13).

The topic 'Respect for Persons' is addressed in light of the Constitution, which recognises and protects fundamental individual rights as well as cultural and communal interests, including protecting the confidentiality of personal information and access to and control over such information. Legally and ethically, people are entitled to make informed choices about their health care and research participation. Especially relevant are privacy interests as genetic information could theoretically be a means to identify a person, even if data are de-identified. The report recommends that the existing legal and ethical framework relating to the protection of personal information, access to, control over and use of personal information and data be revised to provide for a well-aligned and integrated framework that balances a range of diverse interests. The *Protection of Personal Information Act* (*No. 4 of 2013*) (POPI) provides some guidance on how to manage disclosure and sharing of personal information. The panel recommends that 'Engagement with the Information Regulator, the Department of Justice and Constitutional Development, is important to discuss the development of regulations in the POPI Act (*No. 4 of 2013*) and how this will impact on genetics and genomics research' (Recommendation 7). Informed consent is required, as usual, for genetics and genomics work, whether for diagnostic, therapeutic or research purposes. The consensus report endorses the current Department of Health's *Ethics in Health Research Guidelines* (2015) which advocates for broad, tiered or specific consent, according to the study protocol and the choice of the participant or sample provider. The panel recognises, however, that there is lack of consensus regarding the impact of the POPI Act on broad consent and that the situation may change once the Act is fully operational and clarity is obtained from the Regulator. The report recommends that a national consent template for genetics and genomics would be useful so that the essential considerations are not omitted (Recommendation 6).

The third thematic area, 'Good Stewardship', emphasises the inherent characteristics of integrity, honesty, accountability and sharing that inform the notion of stewardship. Policies, legislation, regulations and guidelines must govern genetic and genomic testing; accreditation of laboratories; qualification and certification of staff; and mechanisms and circumstances for feedback of individual results, especially incidental findings. National frameworks for biobanks and access to samples and data are necessary to promote equitable and responsible sharing that will enhance knowledge generation and translational science while aligning with international policies and guidelines (Recommendation 10).

Matters pertaining to ownership/custodianship of tissue, DNA and other biological samples, particularly in the context of secondary use of data and samples, as well as benefit sharing, remain contentious and continue to stimulate debate. The panel suggests that these matters should be considered in particular by the South African Law Reform Commission and the National Intellectual Property Management Office. It is recommended that the country should debate, explore and adapt the recommended 'sociologically informed model' to identify principles of custodianship/ownership of samples and benefit sharing, because the topics have a cascade of implications: ethical values of equity and distributive justice; good governance principles of benefit sharing; and whether intellectual property can exist if genomic resources are to be regarded as a 'common good' (Recommendations 12 and 18).

Given the complexity and rapid evolution of the fields of genetics and genomics, the panel suggests that a 'South African Human Genetics Advisory Board should be established' (Recommendations 11 and 16). The recommendation is that 'The Board should have appropriate expertise to provide guidance to policymakers and regulatory structures'. Furthermore, 'The South African Human Genetics Advisory Board (SAHGB) should be adequately resourced and independent, with the aim of providing oversight in genetics and genomics at the national level and working in concert with ethics and legal regulatory structures'.

One of the most significant overarching recommendations concerns the need to develop capacity in human and material resources. This includes in particular, technical, scientific, computational, bioinformatics and statistical analysis, as well as financial, legal and ethical expertise. Training of genetics nurses, genetics counsellors, medical geneticists, medical scientists, bioinformaticists, biostatisticians and forensic scientists for the public and private sectors is critical to provide a service platform and sustainable genomics programmes to benefit the nation (Recommendation 14). The United Kingdom has committed to their programme of 'Genomics England' within which the concept of workforce training has been firmly embedded (https://www.genomicsengland.co.uk/about-gecip/for-gecip-members/education-and-training/), and this could be a model for the development of genomics in South Africa. The panel recommends that effective measures be implemented to 'improve the public's knowledge and understanding of genetics, genomics and associated new technologies in a culturally sensitive and appropriate manner'; to raise awareness of predatory practices where and when they arise; and to educate the public. The Departments of Basic Education and Higher Education and Training should be encouraged to integrate education about new health-related technologies into primary, secondary and tertiary education curricula. Appropriate genetics and genomics training should be promoted for healthcare professionals, and a substantive investment should be

made into 'training of genetic counsellors and clinical geneticists and other relevant professionals to increase the national capacity to deliver genetics and genomics services' (Recommendation 2). The panel also suggests that courses on forensic DNA testing should be incorporated into the curricula of law degrees.

A strong legal and ethical framework is required that includes: review and oversight roles for research ethics committees and data access committees; engagement with the Information Regulator and the Department of Justice; as well as clear expectations about avoiding harm or offence when reporting research findings (Recommendations 7 and 17). In order to ensure that professionals working in the fields of genetics and genomics adhere to the highest standards of practice, the panel highlights the need for responsibility and a code of conduct (Recommendation 9) regarding sustainable and careful use of genomic resources (reflected as both a value and a practice) by individuals, communities, organisations, companies and governmental institutions. This code of conduct could be established in consultation with professional councils like the Health Professionals Council of South Africa and the Nursing and Pharmacy Council as well as other appropriate professional boards/councils or regulatory bodies, and should be aligned with values and principles from international best practice. It is also necessary to ensure that sanctions for non-compliance with requirements exist and can be enforced. Consequently, authority to impose and enforce sanctions must be allocated appropriately, and the panel recommends that 'Sanctions for non-compliance with current and future legislation must be defined, be implementable and be effective' (Recommendation 19).

Finally, when accessing personal information (demographic, phenotypic and genotypic – which potentially could be used to identify a person), individuals must be protected to avoid social harm such as negative perceptions, stigmatisation and discrimination. In addition, participants must be given the option to gain access to this information, although which information should be made available and to which degree, as well as the way in which information is provided, remain controversial and continue to be debated. Likewise, ownership of data and samples remains a contested issue and researchers must be aware that they are custodians of data and samples, rather than legal owners thereof.

In summary, the consensus report emphasises the benefits to be derived from genetics and genomics work in research, clinical practice and forensic science, and the need for boundaries to be clearly defined and policies adhered to so that the benefits are shared by all while avoiding unnecessary harm. From a genetics and genomics perspective, material differences exist between individuals and between different groups of people. Tools are needed to utilise these differences to better manage prevention, diagnosis and treatment of disease. However, it is increasingly recognised that the analysis of findings for an individual affects not only the immediate biological family members directly (through transmission of heritable traits as genetic information is both personal and familial at the same time), but also indirectly the community to which an individual belongs. Sensitive findings may provoke emotional reactions; negative perceptions about findings can therefore affect large numbers of people, and this should be considered in the regulation of genetics and genomics for the country. Laws and regulations relating to genetics and genomics must be aligned and consistent, and where necessary drafted or updated from time to time to remain abreast of new developments in the field. Ultimately, the practice of genetics and genomics should serve the people of South Africa in the spirit of ubuntu.

Acknowledgements

This consensus study was undertaken under the auspices of the Academy of Science of South Africa, and the panel expresses heartfelt thanks to Prof. Roseanne Diab and Dr Khutso Phalane for all they have done to ensure the successful outcome of the project. We are also grateful to the Department of Science and Technology for the funding provided.

Professor ADM (David) Walker: World-class physicist (1937–2018)

AUTHOR:
Manfred A. Hellberg[1]

AFFILIATION:
[1]School of Chemistry and Physics, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Manfred Hellberg

EMAIL:
hellberg@ukzn.ac.za

HOW TO CITE:
Hellberg MA. Professor ADM (David) Walker: World-class physicist (1937–2018). S Afr J Sci. 2018;114(11/12), Art. #a0298, 2 pages. http://dx.doi.org/10.17159/sajs.2018/a0298

PUBLISHED:
27 November 2018

Professor ADM (David) Walker, one of South Africa's most distinguished scientists, died on 17 September 2018, aged 80. He was Professor of Theoretical Physics at the then University of Natal from 1972 (and Senior Professor from 1989) until he retired in 2002. After retirement, he continued as an active researcher until the end, as Emeritus Professor and Senior Research Associate of the now University of KwaZulu-Natal. After undergoing surgery in June, he unfortunately never really recovered. He will be missed by many, including his colleagues and students, as well as the wider space science community, both nationally and internationally.

After matriculating at Umtata High School, Walker went to Rhodes University, where he came under the influence of two top scientists: Professors JA Gledhill (Physics) and D Burnett (Applied Mathematics). Burnett's work on the kinetic theory of non-uniform gases, published in the 1930s in *Proceedings of the London Mathematical Society,* is still cited eight decades later. And Jack Gledhill, a polymath, is regarded by many as being the father of space physics in South Africa.

After his BSc (Hons) (1959) he was employed as a Junior Lecturer and Lecturer at Rhodes (1960–1962), while doing an MSc on 'Solar eclipses and the ionosphere' with Gledhill. This work led to a joint publication, as well as a single-author paper in *Nature.* Awarded a Shell Postgraduate Scholarship to St. John's College, Cambridge, he did research under Budden (FRS, 1966) at the Cavendish Laboratory. His PhD thesis (1966) entitled 'Radio waves in the ionosphere and exosphere' led to three further single-author papers in leading journals, on the theory of radio propagation. After returning to Rhodes he was promoted to Senior Lecturer, before his move to Durban in 1972.

David Walker was recognised internationally as a world leader in space physics. After his early research on radio propagation and ray tracing, particularly of very-low frequency radio waves, much of his later work centred on ultra-low frequency pulsations of the earth's magnetic field with periods of the order of seconds, hydromagnetic waves in the magnetosphere (the region of space affected by the geomagnetic field), and related magnetospheric and ionospheric phenomena. In addition to carrying out sophisticated mathematical-computational calculations, he was also involved in the analysis and interpretation of both ground-based and satellite observations of the behaviour of the ionised gas (plasma) of geospace.

In 1977–1978 he spent a sabbatical as an Alexander von Humboldt Fellow at the Max Planck Institut für Aeronomie (Lindau-Katlenburg, Germany), where he joined Ray Greenwald. The 1979 paper by Walker, Greenwald, et al. is a classic and has been cited 260 times. Applying Walker's theoretical insight and understanding to the analysis of data from STARE (Scandinavian Twin Auroral Radar Experiment), a concept developed by Greenwald, the paper provided complete understanding of the origin of continuous long-period geomagnetic pulsations (Pc5) arising from hydromagnetic field line resonances. It also showed that STARE-type radar set-ups can provide a powerful new diagnostic technique for geomagnetic pulsation phenomena and hence deepen our understanding of the magnetosphere. In due course, a global network of such dual auroral radars, SuperDARN, was set up.

Walker was a co-initiator of and a Principal Investigator from 1993 to 2002 on the international project known as SHARE (Southern Hemisphere Auroral Radar Experiment) involving dual radars at the South African Antarctic base (SANAE) and the British base (Halley Bay). More generally, he was a highly regarded member of the international SuperDARN community.

When the Foundation for Research Development (FRD), forerunner of the National Research Foundation, introduced international peer evaluation in 1984, he was one of the first group of about a dozen scientists across all disciplines to earn an A-rating (leading international scientist). Despite university, national and international management roles, he retained his A-rating throughout successive re-evaluations to his retirement in 2002, after which it dropped to a B-rating (internationally acclaimed). He was also the first Chair of the FRD Physics Evaluation Committee (1984–1990), and a member of both the Main Awards Committee and the FRD Collegium.

Walker received many accolades. In 1998, the South African Institute of Physics (SAIP) awarded him its highest honour, the biennial SAIP de Beers Gold Medal, and he was elected a Fellow of SAIP when that honour was introduced in 2012. He was also elected a Fellow of the Institute of Physics (London) (1976–1985), the Royal Society of South Africa (1988–) and the University of Natal/KwaZulu-Natal (1989–), and was a Founder Member of the Academy of Science of South Africa (1994–).

For many decades, he played an important role in a range of national committees for radio science, space science and Antarctic research, and represented South Africa internationally on the parent bodies, with acronyms such as URSI, COSPAR, SCOSTEP and SCAR. Amongst others, he served on the Advisory Committee of the Hermanus Magnetic Observatory, chaired the National Committee for the International Polar Year (2006–2008), was on the International Review Panel for the National Space Programme led by SANSA (2012), and was a SANSA Board member in

2013–2014. After his retirement, he served on the NASSP (National Astrophysics and Space Science Programme) Steering Committee and taught a NASSP master's module on 'Magnetohydrodynamic waves in space' for several years. He was also on the Editorial Advisory Boards of both the *South African Journal of Physics* and the *South African Journal of Antarctic Research.*

Walker interacted with many other leaders in the field. Apart from his sabbatical in Germany, his sabbatical and research visits included a year at Cambridge University, three visits of 3–9 months at Johns Hopkins University (Maryland, USA) and 3 months at the British Antarctic Survey (Cambridge, UK). He also gave a 6-week postgraduate course of his choice as Guest Professor at KU Leuven (Belgium).

Internationally, his research and his views were highly regarded. Walker was elected Vice-President of the ICSU Scientific Committee for Antarctic Research (SCAR) in 1998–2002 and Chair of the Solar Terrestrial and Astrophysical Research Working Group of SCAR (1994–2000). He also served on the Editorial Advisory Board of *Planetary and Space Science* (1982–1992), and in 1999 received an Editor's Award for excellence in refereeing from the prime US journal, *Journal of Geophysical Research – Space Physics*, before being appointed an Associate Editor (2000–2002).

Walker published 85 peer-reviewed research articles, many of them singly authored, and his work has been cited more than 3000 times. His latest three papers appeared in 2016, but ongoing projects should lead to further publications. In addition, he presented 79 papers at international and 56 at national conferences. Unusually, for a physicist, he also wrote two major research books: *Plasma Waves in the Magnetosphere* (Springer, 1993) and *Magnetohydrodynamic Waves in Geospace: The Theory of ULF Waves and their Interaction with Energetic Particles in the Solar-Terrestrial Environment* (Institute of Physics Press, 2004). He graduated 8 PhD and 10 MSc students, and a final student submitted his master's thesis recently.

Walker was a consummate all-round academic. Apart from being a world-class researcher, he took teaching seriously. When he joined the University of Natal, he brought a fresh look to our teaching and had a significant impact, for instance, in the development of lecture demonstrations and in supporting teaching experiments. His lectures were known for their clarity, his depth of understanding of physics and his well-planned notes.

During his career, he filled various management roles at the University of Natal, including Head of the Department of Physics for 16 years, Dean of Science, part-time Pro-Vice Principal (Information Systems) for 2 years, Acting Vice-Principal (for three short spells), and member of Council and numerous committees of Senate. He even chaired the Trustees of the University Retirement Fund. After retiring, he held part-time posts as Pro-Vice-Chancellor (Research) (2003–2004) and as Director of Special Programmes in the UKZN Research Office (2005–2006). These involved a variety of projects, including, for instance, merging 3 independent and 16 branch libraries, and setting up a Research Ethics Guide and protocols for the new, merged institution.

His academic leadership style was characterised by his friendliness, his analytical, incisive mind, his integrity and fairness, his ability to delegate and his decisiveness. As a retired colleague wrote: 'It was a pleasure to be a member of the Physics Department he led so superbly – many happy memories!'

Walker had a way with words. The numerous documents that he prepared in his management roles were extremely well written, analytical and to the point. He was also an excellent speaker, whether as a debater in Senate, or as an entertaining after-dinner speaker. He was very well read, and that attribute was invariably reflected in his writing and his speeches.

After completing his second research monograph, Walker embarked on serious historical research, scouring museums and archives, following the story of his forebears, who were 1820 Settlers. This led to a 492-page book entitled *Pawns in a Larger Game: Life on the Eastern Cape Frontier* in 2013.

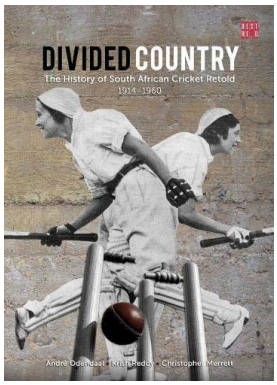
He enjoyed listening to classical music (he was a regular at the KZNPO concerts for many years), gardening and watercolour painting. At one stage he took up long-distance running, and twice qualified for the Comrades Marathon, but to his regret did not finish.

During his Cambridge days, he met Carol Glencross, a Scottish statistician, and they married in 1967. Carol was very supportive of him throughout his career, and his family played an important part in his life. Their three children and six grandchildren now live in New York, Cape Town and Glasgow. Heartfelt condolences go to Carol and the family at this sad time.

A holistic story of South African cricket across time, space, identity, race and gender

BOOK TITLE:
Divided country: The history of South African cricket retold, 1914–1960

COVER:



AUTHORS:
André Odendaal, Krish Reddy and Christopher Merrett

ISBN:
97819282246169

PUBLISHER:
Best Red, HSRC Press, Cape Town; ZAR295

PUBLISHED:
2018

REVIEWER:
Vishnu Padayachee

AFFILIATION:
School of Economic and Business Sciences, University of the Witwatersrand, Johannesburg, South Africa

EMAIL:
vishnu.padayachee@wits.ac.za

HOW TO CITE:
Padayachee V. A holistic story of South African cricket across time, space, identity, race and gender. S Afr J Sci. 2018;114(11/12), Art. #a0293, 2 pages. https://dx.doi.org/10.17159/sajs.2018/a0293

PUBLISHED:
27 November 2018

If the history of South African cricket was to be written, told or retold I cannot think of a better threesome than Andre Odendaal, Krish Reddy and Christopher Merrett to undertake the task. They have each contributed hugely to other projects on the subject of South African cricket in the same spirit as can be found in this book, the second volume in what they expect to be a four-volume narrative. Odendaal is an historian, Reddy a former school principal and Merrett was a university librarian. Each of them brings unique skills, knowledge and experience as academics and cricket historians to this overall project. Among other books, Odendaal wrote the beautifully produced coffee table book *The Story of an African Game: Black Cricketers and the Unmasking of One of Cricket's Greatest Myths, 1850–2003*[1], Reddy co-authored *Blacks in Whites: A Century of Cricket Struggles in KwaZulu-Natal*[2] and Merrett wrote *Caught Behind: Race and Politics in Springbok Cricket*[3].

Here they have pooled their complementary talents to produce a very significant and holistic story of South African cricket across time, space, identity, race and, notably, gender. The role of women in South African cricket would come as a 'total revelation' to most readers, as the authors themselves correctly claim (p. ix). Women's cricket is admirably covered in this book even though the major stories lie in their retelling of the origins and history of the various, sadly racially defined, associations that were formed after the Great War. But the book provides many such revelations, and in this sense is perhaps not a 'retelling' – the story has simply not been told before on this scale, let alone in this detail.

The cricket saga that unfolds happens in the context of the turbulent and divisive history of South Africa; the politics and the cricket story they retell illuminate the same, often inglorious past. These are not two parallel themes but different sides of the same complex reality of modern industrialised South Africa. This second volume picks up the story from the creation of the South African Union, the beginning of the Great War and Gandhi's return to India after some 20 years. It ends (roughly) with the end of the Second Great War, the election of the National Party and the advent of the Afrikaner's own version of liberation and with Gandhi's death by assassination in India. The period has been characterised by Saul Dubow as a time when a 'white South African nation' was being built up, and politics, science, culture and even sport were among the arenas in which this fragile post-South African War (or Boer War) project was played out. For example, Dubow shows how 'science was used to underpin a sense of South African patriot achievement within the broad context of Commonwealth [and imperial] belonging'[4(p.13)]. Cricket and rugby played this 'white nation-building' role in the field of sport.

Earlier attempts at forging non-racial cricket in the form of the Barnato tournaments that began around 1888 collapsed in the first half of the 20th century, through the selfish and self-centred desire by some among the various 'race groups' to go their own way. Turning their backs on the nascent non-racial cricket that was being played for the Barnato Trophy, separate racially based leagues were formed for coloured (1926), Indian (1940) and 'Bantu' (1932) players, and later a Malay organisation split from the coloured league which was perceived as being dominated by Christian coloured players. A white women's organisation was formed in 1952 that also took in the organised game in then Rhodesia. These developments mirror to some extent what happened in South African professional football which went 'multinational' in the 1970s and 1980s.[5]

None of these cricket organisations or their mind-blowing acronyms has survived to today. All the while white South African cricket – the 'institutionalised representative of apartheid in cricket' (p.147) – thrived under state support and protection and was spreading its wings internationally, albeit limited to playing their cousins in the British Empire (England, Australia and New Zealand).

These rather ironic developments in local black cricket only began to change in the 1950s. The story of these racially based cricket organisations, their origins and development may at times fill readers with a real sense of revulsion, yet telling their story, however unsavoury, with wholeness and proportionality is vitally important. Among other things it allows the reader to appreciate that cricket among black people in South Africa has its own rich history, both of resistance and accommodation to white rule. It gives the lie to the view expressed by former Springbok rugby captain Dawie de Villiers that 'blacks have only really known Western sport for the last ten years' (p.9).

The cricketing history of each of these leagues is told in great depth and with great care and sensitivity. The book consists of five parts and 32 chapters. The first two parts cover the general narrative of South African cricket set in the context of the unfolding racial saga and identity politics that was and still is South African politics. Parts 3 and 4 provide an overview of each of these racially separate cricket organisations. In Part 5, award-winning cricket historian and statistician Krish Reddy fills in the missing tale of their records and scores with meticulous care. This has to be an (unrewarded) labour of love as well as one requiring a detective's skill. Reddy managed to track down some of the information, photographs and records contained in this book in the basement of the Curries Fountain sports stadium. It is worth buying the book if only for the photographs and other visual material contained in its 442 pages. (Reddy was awarded the International Cricket Council's Volunteer's Medal in December 2009 in recognition of his 'outstanding service to cricket' and the UK-based Association of Cricket Statisticians and Historians chose him as their Statistician of the Year in 2007.)

We learn for example of a South African Indian cricket and football team that toured India in 1922, beating the arrival of Clive Rice's widely celebrated mainly white team in Indian by nearly 70 years! We learn that black women – including Mrs Wauchope, the wife of the respected Eastern Cape notable Isaac Williams Wauchope – were an

integral part of cricket's development from the very beginning, when they played a prominent role in hosting the visiting Kimberly team in Port Elizabeth in 1888, at which match local black people arrived as spectators in large numbers. We learn too of the tensions among white cricketers, especially between English- and Afrikaans-speaking players, and of the challenges faced by working class white and Jewish players in breaking into the game, and a lot more.

With three authors involved in its making, the book does have a certain stylistic inconsistency. I much preferred the first two narrative parts and enjoyed less the story of the racially separate leagues, especially their treatment in Part 4. The detail provided in these middle chapters is no doubt essential and important, but it is plainly less enthralling for the general reader.

I recommend the book to anyone interested in the history or politics of cricket, sports history or South African social history, whether in South Africa or internationally.

References

1. Alegi P. Laduma, soccer, politics and society in South Africa. Pietermaritzburg: UKZN Press; 2004.
2. Desai A, Padayachee V, Reddy K, Vahed G. Blacks in whites: A century of cricket struggles in KwaZulu-Natal. Pietermaritzburg: UKZN Press; 2002.
3. Dubow S. A commonwealth of knowledge, science, sensibility and white South Africa, 1820–2000. New York: Oxford University Press; 2006.
4. Merrett C, Murray B. Caught behind: Race and politics in Springbok cricket. Johannesburg: Wits University Press; 2004.
5. Odendaal A. The story of an African game: Black cricketers and the unmasking of one of cricket's greatest myths, 1850–2003. Cape Town: David Philip and HSRC Press; 2003.

Reducing inequality and carbon emissions: Innovation of developmental pathways

AUTHOR:
Harald Winkler[1]

AFFILIATION:
[1]Energy Research Centre, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Harald Winkler

EMAIL:
harald.winkler@uct.ac.za

KEYWORDS:
inequality; climate change; mitigation; development; sustainability

HOW TO CITE:
Winkler H. Reducing inequality and carbon emissions: Innovation of developmental pathways. S Afr J Sci. 2018;114(11/12), Art. #a0294, 7 pages. https://dx.doi.org/10.17159/sajs.2018/a0294

PUBLISHED:
27 November 2018

Professor Harald Winkler is the recipient of the 2017/2018 NSTF-South32 Special Annual Theme Award: Sustainable Energy for All (in recognition of the United Nations 'International Decade of Sustainable Energy for All').

Dual challenges of inequality and mitigation

Inequality and poverty are the top priorities in South Africa's National Development Plan[1]; job creation and education are key means to reduce both. At the same time, the country wants to make a fair contribution to global efforts to combat climate change. In 2015, the global community adopted both Sustainable Development Goals (SDGs)[2] and the Paris Agreement[3]. To understand what it is to be human in the 21st century, and particularly in South Africa, one needs to consider high inequality[2,4] and dangerous climate change[3,5]. Beyond analysis, the challenge is to reduce inequality and greenhouse gas (GHG) emissions – which is the motivation for this article. To achieve that, innovative pathways to development will have to be charted. In this respect, local is the new global and we 'need to bury the notion that global is not African'[6]. This article starts with a South African focus, considered integral to global challenges of inequality and mitigation.

Inequality and mitigation, locally and globally

Consider inequality in South Africa. Inequality has many dimensions, but while Thomas Pikkety argues compellingly that asset inequality is more persistent than income inequality[7], the latter is the more common measure, including in South Africa. Figure 1a illustrates a notional household of five people – one might think that with a monthly income of ZAR50 000, this household would be solidly in the middle of the South African distribution. However, the actual position is the green line, whilst the median value is shown by the small red line. This observation is underpinned by the robust overall finding of a review of the economics of income inequality: 'that inequality in incomes is extremely high from a global comparative perspective and has increased since the democratic transition in 1994'[8].

Income inequality is a persistent feature globally. While the world is no longer divided into two groups of developed and developing countries, it is also not homogeneous (see http://www.ecomagination.com/hans-rosling-and-the-future-of-the-world; or watch Rosling's brilliant talk at http://www.ted.com/talks/hans_rosling_and_the_magic_washing_machine.html). Figure 1b shows a spectrum of income distributions by ventile, or twentieth of the population. Milanovic[9] shows that the poorest 5% of Americans have the same income as the richest 5% of Indians – about USD3000–4000 per month.

Further research is needed into such distributions by wealth or assets, and more work is needed to provide a view of inequality of GHG emissions across households. The trade-off between affluence and household emissions has been analysed for Indonesia,[10] with similar work being initiated in South Africa. At this stage, we cannot show comparable graphs of inequality by GHG emissions across income groups. What we do know is that inequality in access to energy services is a critical factor in (South) Africa and across countries.[11]

The inequality of global energy CO_2 emissions compared to poverty are illustrated in Figure 2. Country areas are adjusted for cumulative emissions, since 1850 in Figure 2a. A radically different map is generated in Figure 2b, which maps current levels of poverty.

It is a deep injustice that those less responsible for the problem of climate change are most vulnerable to its impacts. Not only do poor countries and communities have lower capacity to adapt, or recover from loss and damage, but they are expected to take on some burden of reducing emissions in future.[13,14] What is needed is not only zero poverty, zero carbon – but also zero impacts. The coping capacity across African countries is a major concern.[15]

There is a very wide range of possible interactions between scholars who investigate poverty and inequality and the climate change community of practice. The focus in this commentary is on mitigation within climate change (distinct from physical science, and impacts, vulnerability and adaptation), and on inequality as a focus that sharpens the focus on poverty.

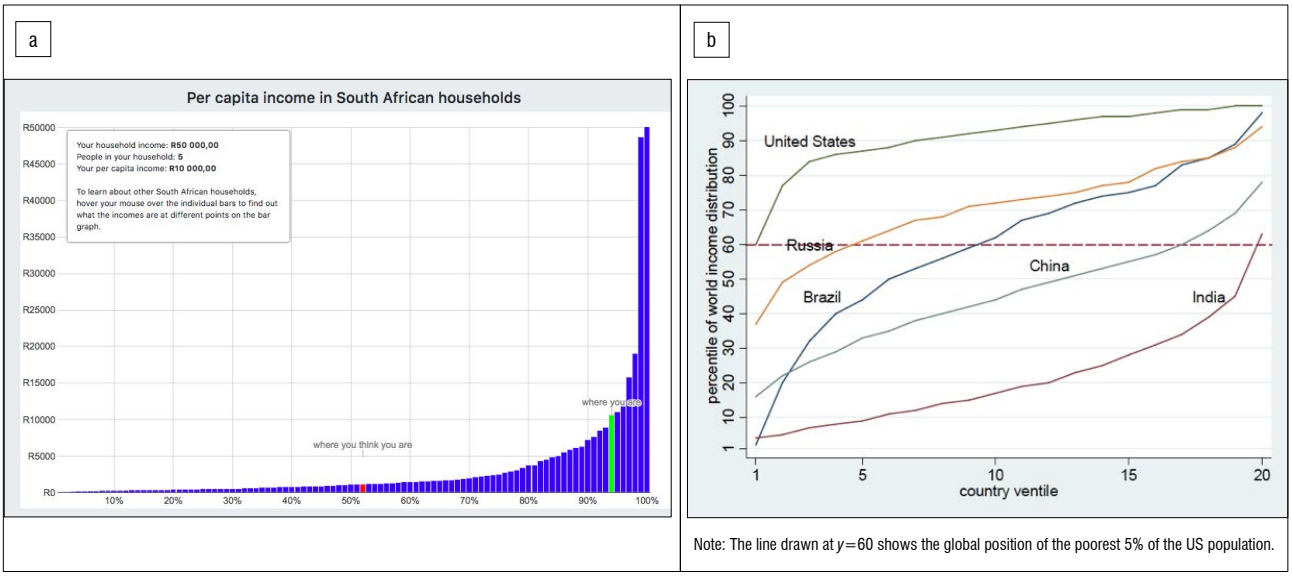
The focus on mitigation is partly because it is the author's research interest, but more fundamentally because GHG emissions are the root cause of climate change. After 17 years of assessment, the Intergovernmental Panel on Climate Change found that warming of the climate system is 'unequivocal' and '*very likely* due to anthropogenic GHG emissions'[16]. Reducing GHG emissions goes to the root of the problem, and in that sense is a radical solution. To address both inequality and mitigation is to ask some fundamental questions.

The challenges of development and climate are being considered globally. Figure 3 shows the SDGs, highlighting Goal 1 to 'end poverty in all its forms everywhere', Goal 10 to 'reduce inequality within and among countries', and Goal 13 to 'take urgent action to combat climate change and its impacts'[2]. The triangle linking the three goals is the focus of this article.

The Paris Agreement represents the best, if imperfect, efforts of the global community to respond to climate change. It is considered a hybrid architecture, including bottom-up and top-down elements.[17,18] Probably the most significantly

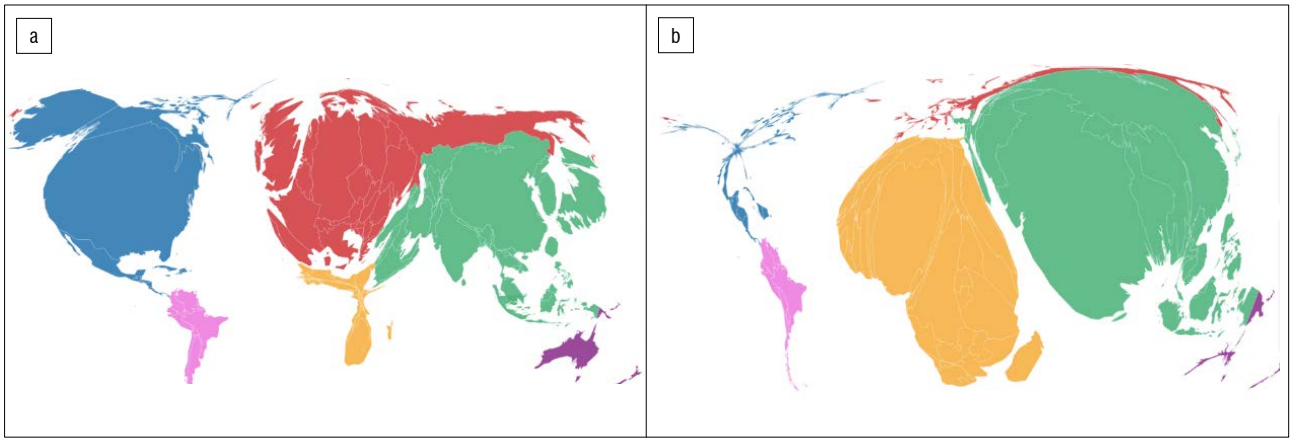
new elements are nationally determined contributions (NDCs), with countries deciding what to commit to rather than establishing that in a multi-lateral negotiation. Top-down elements include goals relating to temperature (in Article 2.1a), mitigation (Article 4.1), adaptation (Articles 2.1b and 7.1) and finance (Article 2.1c and paragraph 53 of the Paris decision[19]), as well as mandatory review (Article 13) and a global stock-take (Article 14). As of October 2018, 177 countries had submitted their first NDCs, demonstrating near-universal participation in mitigation (consult the NDC registry at http://www4.unfccc.int/ndcregistry/Pages/Home.aspx). This is in contrast to the Kyoto Protocol, which in 1997 negotiated strong mitigation commitments only for developed countries. The scale and intensity of the challenge requires developing countries – even though they have contributed less to the problem[14] – to also contribute to the solution. Developed countries will need to rethink their development paradigm. Collectively, the sum of NDCs puts us on a path towards 2.7–3.1 °C, although much depends on what countries do after 2030.[20]

Achieving national development goals, the SDGs and contributing to the Paris Agreement will require innovative development paths – which should be informed by long-term GHG *development* strategies (Article 4.19). South Africa's NDC contains a mitigation component with absolute numbers – keeping GHG emissions between 398 and 614 Mt CO_2-eq in 2025 and 2030, to be achieved in the context of development. The challenges are particularly sharp in South Africa, which with its persistently high Gini coefficient and coal-based energy economy can be seen as a litmus test for addressing both inequality and mitigation.



Sources: (a) SALDRU income comparison tool[67]; (b) Milanovic[9]

Figure 1: Income inequality in South Africa and across selected countries by ventile.



Source: The carbon map[12]

Note: (a) Country sizes show CO_2 emissions from energy use, 1850–2011. These historical (or 'cumulative') emissions remain relevant because CO_2 can remain in the air for centuries. Europe and the USA dominated, having released about half the CO_2 ever emitted.[12]

Note: (b) Country sizes show the number of people living on less than USD1.25 a day. Poverty adds to climate change vulnerability because lack of access to health services increases the risks of climatic changes, and lack of access to capital makes it harder to implement adaptation measures. Population in 2013, share for most recent year.[12]

Figure 2: World maps showing areas of countries adjusted by (a) cumulative energy CO_2 emissions and (b) poverty.

Source: United Nations[2]

Figure 3: Sustainable development goals on poverty, inequality and climate change; links between goals added.

Development pathways that reduce inequality and GHG emissions

How to reduce inequality and GHG emissions? This question frames the overall outcome that is required in South Africa and the world. It is deliberately posed simply, and suggests a state-of-the-world outcome. Answers to this question are well beyond the control of any single actor or institution, never mind any academic. Yet the overall question is useful to keep in mind in addressing more precise research questions.

Research question

How could innovative development pathways reduce inequality and GHG emissions? This question will require a long time and many minds to address. As any good long-term research question should, it raises further questions. What innovation is needed to follow such development pathways?[21] How do GHG emissions and inequality correlate in South Africa, considering multiple dimensions of inequality and drivers of GHG emissions? How does that compare to other countries? How do we need to think differently, to shift to new development pathways that both reduce inequality and enhance mitigation? What are the implications for systems, policy, technology, investment, goals and mind-sets? Given that past patterns of development have 'baked in' high emissions into existing energy infrastructure[22] – how do we avoid repeating the mistake?

Approach

How do researchers best think about normative issues, such as procedural and distributional equity, which are integral to inequality and mitigation? The approach in this article is to aim at rigorous analysis, but not pretend to be value free. Good analysis must be based on best available data and replicable methodologies, seeking to be as systematic as possible. As a community of scholars, we continually must check for confirmation bias, and remain open to results that we do not expect or like. But rather than pretending to know a universal truth, we do better by stating upfront the values we hold and any conscious biases that may influence the analysis. This author has made clear that key goals should be zero carbon and reduced inequality[23] – and zero poverty and zero impacts. These are matters of investigation, as well as important goals to adopt. Having said that, the remainder of this section sets out means for rigorous analysis.

Towards a theoretical framework

Research on inequality and mitigation must draw on multiple disciplines – within each community of practice and *a fortiori* across them. Energy research has no single theory and analysis of mitigation draws on several disciplines – political studies, economics, social sciences, engineering and more. Scholars investigating inequality similarly come from a range of disciplines – perhaps most often from economics, but also sociology and other social sciences. Research on inequality and mitigation is necessarily interdisciplinary, and in Winskel's terms, requires not just cognate but radical interdisciplinarity.[24] Furthermore, co-production of knowledge with a range of stakeholders increases the influence of research dramatically,[25,26] so that transdisciplinarity becomes essential. There are no existing theories for such research. Constructing theoretical frameworks which borrow, rather eclectically, from a range of theories, is both a strength and a weakness. The weakness lies in not having a unifying explanatory theory, which can lead to lazy thinking. The strength is that the diversity of theoretical approaches ensures creative tensions, conducive to innovation and quite capable of systematic arrangement. A theoretical framework for inequality and mitigation is an important objective.

The real-world challenges of inequality and mitigation in the 21st century require systems thinking. The leverage points to intervene in systems were compellingly outlined by Meadows, who went on to warn that it would be a terrible mistake to assume that 'here at last, is the key to prediction and control'[27] and to advise to remain humble[28]. The pedagogy of inequality and mitigation should be iterative, drawing on Freire's action-reflection-action cycles.[29] An updated version of continuous learning and adjustment might be called adaptive management, which is relevant to both inequality and climate change.[30]

Quantitative and qualitative analyses

Addressing the research question will require complex problem-solving using both quantitative and qualitative analyses. The existing literature on mitigation has historically been based on techno-economic modelling.[31,32] Broadly, three areas for reducing GHG emissions from energy use and supply have been identified: (1) improving energy efficiency, (2) changing the fuel mix to lower carbon sources and (3) moving to less energy- and emissions-intensive sectors of the economy.[33] Careful modelling and analysis provides rigour in complex problem-solving, has inherent value and will continue to provide a counter-balance for hand-waving analysis of mitigation scenarios – or indeed development pathways.

To address inequality and mitigation, correlations will be important. Learning from research on Indonesia[10] and internationally[34], we need to understand how GHG emissions and inequality correlate in South Africa. Some methodologies that are applicable to inequality *within* countries include the use of input-output tables[10] or social accounting matrices, in order to attribute all emissions within a country to households; and decomposition analysis of emissions growth drawing on the Kaya identity[35]. Irfany and Klasen[10] found that 'aggregate consumption is the most important driver of carbon footprints', so having consumption-based GHG inventories for South Africa will be an important research task. It is important to distinguish direct GHG emissions in households, indirect GHG emission elsewhere in the economy (e.g. energy, steel, cement) and those embodied in trade.[36] Further extensions should consider concentration of assets (as distinct from income), urban–rural differences or spectrums, differences in locational value because of different places of production or consumption[37], inequalities in skill levels, direct and indirect energy use[38], and other potentially significant contextual drivers. In many African countries, although less so in South Africa, the problem is one of avoiding emissions, rather than reducing from high levels.[39] Comparative analysis between countries would be the next step (of inequality *within* each) – and points to the need for an international team of researchers.

Research on inequality *among* countries (as in SDG 10) would likely draw on metrics such as the Gini coefficient[40] (and Lorenz curves), the Theil Index[41], and an extensive literature on environmental Kuznets curves, including application to climate change[42]. The Gini coefficient has been applied to carbon, with the finding that '70% of carbon space in the atmosphere has been used for unequal distribution, which is almost the same as that of incomes in a country with the biggest gap between the rich and the poor in the world'[43]. Some initial global modelling of inequality and mitigation suggests that there may not be only trade-offs, finding that 'aggressive inequality reduction…would realistically increase GHG emissions by less than 8%' over several decades; however, overall reductions are required rather than limiting increases, and the authors point to the need to 'deeper under-explored linkages and synergies between reducing income inequality and climate change'[34]. Another perspective is that 1.5 °C and SDGs can remain within reach with 'low energy demand scenarios'[44] globally; yet how this plays out in poor communities and countries remains important.

Energy is only one – albeit an important – input for development. Research needs to understand inequalities in energy use and supply. Sustainable energy for all is a critical challenge in South Africa, Africa, other developing countries and the world in the 21st century. Much energy analysis tends to focus on GHG emissions associated with energy production – the use of coal, oil and gas to generate electricity and supply liquid fuel.[45,46] Yet patterns of consumption are important to understand inequalities. Energy analysis of household consumption is critical in this regard and analysis of changing patterns in India[38] is highly relevant to South Africa. Addressing highly unequal access to affordable modern energy services in many developing countries[47] is a key input to development. The question is how this can be done in a low emissions manner.

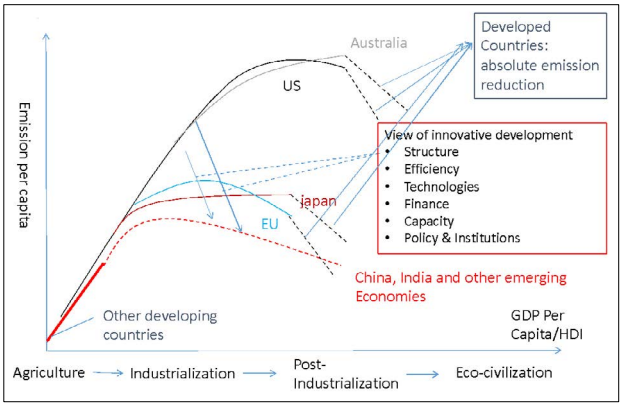
Renewable energy for electricity is one rapidly growing system that promises synergies. Only a few years ago, the assumption was that renewable energy technologies imposed an incremental cost, being more expensive than relatively cheaper fossil fuels. With very rapidly falling costs, there is now a net benefit – globally[48], in sub-Saharan Africa[49] and South Africa[50]. Already coal communities are being impacted by mine closures, with other mining sectors increasingly being automated. Will workers from those communities find employment in emerging sectors, including energy service companies and renewable energy? A just energy transition is key to development pathways to provide jobs and livelihoods for the future.

Yet existing energy systems in developing countries are not able to make transformative inputs to achieving development outcomes for the well-being of the majority of people. Hundreds of millions of people in developing countries still lack access to affordable energy services.[47] On current trends, it will take until 2080 to reach universal access to electricity across the African continent[51] – and move beyond fuel from solid biomass (essentially fuelwood and charcoal), on which 4 out of 5 African households depend[52].

Development pathways

Development pathways, as distinct from mitigation pathways, are key to our (future) analytical frameworks. What development pathways would meet South Africa and Africa's development objectives? What storylines reduce poverty, inequality and GHG emissions?

Figure 4 offers one conceptual framing to think about such questions. If countries followed the innovative development pathways presented in Figure 4, would this reduce the patterns of income inequality shown earlier in Figure 1?



Source: Zou et al.[21]

Figure 4: A perspective on global innovation of development pathways.

The challenges of inequality and climate change mitigation in the context of development are each 'wicked problems'; combining them constitutes a super-wicked problem.[53] Indeed, the inequality-mitigation nexus as a super-wicked problem poses fundamental questions about the future of industrial civilization.[54] In our country, the extreme degrees of inequality threaten to undermine prospects of a better life. Globally, the high-emission development paths which developed countries have followed cannot be the model for the future.[14,55-57] What is needed is to change from emissions intensive development paths[34] to innovative development pathways[21] that reduce emissions, poverty, inequality and emissions. The research agenda should consider *'living well' with less* – so scenarios of lower demand globally[44] are helpful to our analysis – but cannot come at the expense of those in energy poverty.

Is our education system training young people with the appropriate skills for a major transition? Three key skills that will be needed in future are 'complex problem-solving, critical thinking and creativity'[58]. Will there be kinds of work that are irreducibly human? The future will likely be very different from our past and present, so understanding change is crucial to unpacking inequality and mitigation.

Understanding change

How do we change development pathways?[33] It seems safe to assume that no single actor is in charge of development. No single government, company, union, investor, social movement, city or other agent of change on their own changes development pathways – these are the result of myriad decisions by large numbers of actors. To address this research question, understanding of change agents (those preceding and others), determinants of change, and adaptive management, among other issues, is required. Are key determinants of change to be found in material conditions, ideas, institutions[59,60], or networks, recalling that the mind-set or paradigm is a high-level lever to intervene in a system?[27] How must we think differently, in order to shift to new development pathways that reduce inequality and GHG emissions?

How must policy change and what are specific policy instruments that can reduce inequality and GHG emissions? While policy is no

silver bullet, and needs to be understood as part of systems, policy analysis is an important area requiring more research. Research should examine policy instruments ranging from a universal basic income to pricing with pro-poor revenue recycling,[61] investments in education, a more progressive tax system, a tax on financial speculation, and more. Bearing in mind the future of work, a 'tax on robots' is another instrument requiring close attention. It will be important to pilot, demonstrate and replicate specific instruments; to learn from both successes and failures, and to adapt as we learn by reflecting on action.

Technology is changing very rapidly. Artificial intelligence may be a key component of what the founder of the World Economic Forum thinks may be a 'fourth industrial revolution'[62], while others talk about 'post-capitalism'[63]. Regardless of framing, addressing the super-wicked problem of inequality and mitigation requires thinking about the future – of work, capital, labour and society in general. With the rise of artificial intelligence and biotechnology, much of humanity may become militarily and economically useless[64], with much of today's paid labour replaced by machines. In a 'post-work' future, the contest may not be about capital and labour, but around energy and resources.[65] This process may further entrench inequality, especially in emerging economies with low skill levels. The patterns of investment will have to shift dramatically from those of the past.

Pursuing a focus on inequality and mitigation within the broader fields of development and climate change should also attract more black South African scholars. The climate community of practice in South Africa is still largely composed of older white men (including the present author). Transformation is essential, and linking mitigation to poverty and inequality can be expected to ground the analysis in issues of interest to emerging scholars.

Managing change to address super-wicked problems in a complex world requires adaptive management. O'Brien and Selboe[30] argue that, beyond technical problems, management of complex adaptive systems requires changes in people's mind-sets, shedding entrenched ways of thinking, tolerating disequilibria – and considering change other than as a linear pathway. Adaptive management is needed to deal with dynamic, social, human and emergent complexity – all of which characterise inequality and mitigation. A long-term time frame is needed, but not a plan from here to there. It is all about planning, not a plan; adapting as you go along rather than rigidly implementing. An objective of the agenda proposed here must be applying adaptive management to follow innovative development pathways, from local to global scales, and addressing long-term problems that require urgent action.

Beyond management, the challenges of inequality and mitigation will require a new social contract.[66] Such a social contract accepts that the poor have to be lifted out of poverty, with little impact on emissions; that richer households can be happier with less; and that the aspirations of middle classes should shift from having more to living well.

Conclusion

Reducing inequality and GHG emissions are key challenges of the 21st century. An agenda for research and co-production of knowledge, framed by the question of how innovative development pathways could reduce inequality and GHG emissions, has been proposed here.

The agenda starts at the national level, to pursue multiple development goals – notably reducing inequality and poverty, and finding low emissions paths to achieve those objectives. It is by action at local scale that we must aim to achieve the SDGs and Paris Agreement globally.

The complex nature of both inequality and mitigation requires interdisciplinary research and a theoretical framework will necessarily draw on multiple theories. The article calls on communities of scholars to work towards a theoretical framework to understand development paths that reduce inequality and mitigation.

The theoretical framework will require a very wide range of methods, both qualitative and quantitative. Engaged scholars of inequality and mitigation will need expertise in consumption-based accounting as much as policy analysis, to name just two examples. Complex problem-solving, critical thinking and creativity will be key skills in thinking about points to intervene in complex systems. The research must be rigorous while declaring its goals to reduce inequality and emissions upfront. While excellent research is an essential foundation and an international collaboration is proposed, co-production of knowledge makes a bigger difference. Learning by doing, together with a range of stakeholders, in a continual action-reflection-action cycle is an essential pedagogy.

Meeting the challenges of inequality and mitigation requires understanding how change happens – change of policy, technology, investment, but perhaps more fundamentally, some determinants and agents of change. This information will be valuable for adaptive management of complex systems. Innovating in development pathways that will shift countries from development pathways from emissions-intensive to low emissions development pathways, while reducing inequality within and across countries – those are certainly complex adaptive systems.

What is required is no less than a new social contract, in which the rich live better with less, the poor are lifted out of poverty, and middle-class aspirations shift from having more to living well. The vision of a world with less inequality and fewer emissions is a bold but necessary one, and achieving it is a challenge worthy of concerted action.

References

1. South African National Planning Commission. Our future – make it work: National development plan 2030. Pretoria: The Presidency; 2012 [cited 2012 Aug 23]. Available from: http://www.npconline.co.za/pebble.asp?relid=757
2. United Nations. Transforming our world: The 2030 agenda for sustainable development. A/RES/70/1: Sustainable Development Goals. New York: United Nations; 2015. Available from: https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf
3. United Nations Framework Convention on Climate Change (UNFCCC). Paris Agreement. Annex to decision 1/CP.21 document FCCC/CP/2015/10/Add.1. Paris: UNFCCC; 2015. Available from: http://unfccc.int/resource/docs/2015/cop21/eng/10a01.pdf#page=2
4. Stiglitz JE. The great divide: Unequal societies and what we can do about them. New York: Norton; 2015.
5. Intergovernmental Panel on Climate Change (IPCC). Climate change 2014: Synthesis report: Fifth assessment report of the Intergovernmental Panel on Climate Change. Geneva: IPCC; 2014. Available from: http://www.ipcc.ch/pdf/assessment-report/ar5/syr/SYR_AR5_SPM.pdf
6. Phakeng M. Remarks in presentation in selection process for Vice Chancellor. Cited in: Farber T. Strong, qualified and able: Two women battle it out for UCT top post. TimesLive 2018 February 22. Available from: https://www.timeslive.co.za/news/south-africa/2018-02-22-strong-qualified-and-able-two-women-battle-it-out-for-uct-top-post/
7. Pikkety T. Capitalism in the twenty-first century. Paris: Éditions du Seuil and Harvard University Press; 2013.
8. Leibbrandt M, Ranchhod V. A review of the economics of income inequality literature in the South African context [document on the Internet]. c2017 [cited 2018 Oct 29]. Available from: http://www.psppd.org/wp-content/uploads/2017/08/Inequality-in-SA_A-review-of-the-economics-of-income-inequality-literature-in-the-SA-context.pdf
9. Milanovic B. Global income inequality by the numbers: In history and now — An overview. Policy Research Working Paper 6259. Washington: Development Research Group, Poverty and Inequality Team; 2012. Available from: http://documents.worldbank.org/curated/en/959251468176687085/pdf/wps6259.pdf
10. Irfany MI, Klasen S. Affluence and emission tradeoffs: Evidence from Indonesian households' carbon footprint. Environ Develop Econ. 2017;22:546–570. https://doi.org/10.1017/S1355770X17000262
11. Tait L. Towards a multidimensional framework for measuring household energy access: Application to South Africa. Energy Sustain Develop. 2017;38:1–9. https://doi.org/10.1016/j.esd.2017.01.007
12. Clark D, Houston R. The carbon map. No date [cited 2018 Oct 29]. Available from: http://www.carbonmap.org/#Historical

13. Shue H. Global environment and international inequality. Int Affairs. 1999;75:531–544. https://doi.org/10.1111/1468-2346.00092

14. Agarwal A, Narain S. Global warming in an unequal world: A case of environmental colonialism. New Delhi: Centre for Science and Environment; 1991. Available from: http://www.indiaenvironmentportal.org.in/files/GlobalWarming%20Book.pdf

15. Sokona Y, Denton F. Climate change impacts: Can Africa cope with the challenges? Clim Policy. 2001;1:117–123. https://doi.org/10.3763/cpol.2001.0110

16. Intergovernmental Panel on Climate Change (IPCC). Climate change 2007: Synthesis report: Fourth assessment report of the Intergovernmental Panel on Climate Change. Geneva: IPCC; 2007.

17. Oberthür S, Bodle R. Legal form and nature of the Paris Outcome. Clim Law. 2016;6:40–57.

18. Klein D, Carazo P, Bulmer J, Doelle M, Higham A, editors. The Paris Agreement on climate change: Analysis and commentary. Oxford: Oxford University Press; 2017.

19. United Nations Framework Convention on Climate Change (UNFCCC). Decision 1/CP.21, document FCCC/CP/2015/10/Add.1 Paris: UNFCCC; 2015. Available from: http://unfccc.int/resource/docs/2015/cop21/eng/10a01.pdf

20. Rogelj J, Den Elzen MGJ, Höhne N, Fransen T, Fekete H, Winkler H, et al. Paris Agreement climate proposals need boost to keep warming well below 2°C. Nature. 2016;534:631–639. https://doi.org/10.1038/nature18307

21. Zou J, Fu S, Liu Q, Ward J, Ritz R, Jiang K, et al. Pursuing an innovative development pathway: Understanding China's NDC [document on the Internet]. c2016 [cited 2018 Oct 29]. Available from: http://documents.worldbank.org/curated/en/312771480392483509/pdf/110555-WP-FINAL-PMR-China-Country-Paper-Digital-v1-PUBLIC-ABSTRACT-SENT.pdf

22. Davis SJ, Caldeira K, Matthews D. Future CO_2 emissions and climate change from existing energy infrastructure. Science. 2010;329:1330–1333. https://doi.org/10.1126/science.1188566

23. Schultz D. Low carbon energy economy with socioeconomic benefits. Mail & Guardian. 2018 June 29. Available from: https://mg.co.za/article/2018-06-29-00-low-carbon-energy-economy-with-socioeconomic-benefits

24. Winskel M. The pursuit of interdisciplinary whole systems energy research: Insights from the UK Energy Research Centre. Energy Res Social Sci. 2018;37:74–84. https://doi.org/10.1016/j.erss.2017.09.012

25. Orford M, Raubenheimer S, Kantor B. Climate change and the Kyoto Protocol's Clean Development Mechanism: SouthSouthNorth – stories from the developing world. Cape Town: Double Storey; 2004.

26. Boulle M, Torres Gunfaus M, Kane L, Du Toit M, Winkler H, Raubenheimer S. MAPS approach: Learning and doing in the global South [document on the Internet]. c2015 [cited 2018 Oct 29]. Available from: http://www.mapsprogramme.org/wp-content/uploads/The-MAPS-DNA_02-07-2015_Final-.pdf

27. Meadows D. Leverage points: Places to intervene in the system. Hartland VT: The Sustainability Institute; 1999.

28. Meadows D. Dancing with systems. Whole Earth Winter. 2001:58–63.

29. Freire P. Pedagogy of the oppressed. London: Continuum; 1970.

30. O'Brien K, Selboe E. The adaptive challenge of climate change. Cambridge: Cambridge University Press; 2015. https://doi.org/10.1017/CBO9781139149389

31. Merven B, Arndt C, Winkler H. The development of a linked modelling framework for analysing socio-economic impacts of energy and climate policies in South Africa. UNU WIDER working paper 2017/40 [document on the Internet]. c2017 [cited 2018 Oct 29]. Available from: https://www.wider.unu.edu/sites/default/files/wp2017-40.pdf

32. Pye S, Bataille C. Improving deep decarbonization modelling capacity for developed and developing country contexts. Clim Policy. 2016;16:S27–S46. https://doi.org/10.1080/14693062.2016.1173004

33. Winkler H, Marquard A. Changing development paths: From an energy-intensive to low-carbon economy in South Africa. Clim Develop. 2009;1:47–65. https://doi.org/10.3763/cdev.2009.0003

34. Rao ND, Min J. Less global inequality can improve climate outcomes. WIREs Clim Change. 2018;9, e513, 6 pages. https://doi.org/10.1002/wcc.513

35. Kaya Y, Yokobiri K. Environment, energy and economy: Strategies for sustainability. Tokyo: United Nations University Press; 1997.

36. Davis SJ, Caldeira K. Consumption-based accounting of CO_2 emissions. Proc Natl Acad Sci USA. 2010;107:5687–5692. https://doi.org/10.1073/pnas.0906974107

37. Milanovic B. The haves and the have-nots: A brief and idiosyncratic history of global inequality. New York: Basic Books; 2011.

38. Pachauri S. An energy analysis of household consumption: Changing patterns of direct and indirect use in India. Dordrecht: Springer; 2007.

39. Davidson O, Halsnaes K, Huq S, Kok M, Metz B, Sokona Y, et al. The development and climate nexus: The case of sub-Saharan Africa. Clim Policy. 2003;3:S97–S113. https://doi.org/10.1016/j.clipol.2003.10.007

40. Yitzhaki S. Relative deprivation and the Gini coefficient. Quart J Econ. 1979;93:321–324. https://doi.org/10.2307/1883197

41. Conceição P, Galbraith JK. Constructing long and dense time-series of inequality using the Theil Index. East Econ J. 2000;26:61–74.

42. Intergovernmental Panel on Climate Change (IPCC). Climate change 2014: Mitigation of climate change. IPCC Working Group III contribution to the fifth assessment report. Geneva: IPCC; 2014. Available from: http://www.ipcc.ch/report/ar5/wg3/

43. Teng F, He J, Pan X, Zhang C. Metric of carbon equity: Carbon Gini Index based on historical cumulative emission per capita. Adv Clim Change Res. 2016;2:134–140. https://doi.org/10.3724/SP.J.1248.2011.00134

44. Grubler A, Wilson C, Bento N, Boza-Kiss B, Krey V, McCollum DL, et al. A low energy demand scenario for meeting the 1.5 °C target and sustainable development goals without negative emission technologies. Nature Energy. 2018;3:515–527. https://doi.org/10.1038/s41560-018-0172-6

45. Green F, Denniss R. Cutting with both arms of the scissors: The economic and political case for restrictive supply-side climate policies. Clim Change. 2018;150(1–2):73–87. https://doi.org/10.1007/s10584-018-2162-x

46. Johannson TB, Williams RH, Ishitani H, Edmonds J. Options for reducing CO2 emissions from the energy supply sector. Energy Policy. 1996;24:985–1003. https://doi.org/10.1016/S0301-4215(96)80362-4

47. Winkler H, Simões AF, La Rovere EL, Alam M, Rahman A, Mwakasonda S. Access and affordability of electricity in developing countries. World Develop. 2011;39:1037–1050. https://doi.org/10.1016/j.worlddev.2010.02.021

48. International Renewable Energy Agency, International Energy Agency, Renewable Energy Policy Network for the 21st Century (IRENA, IEA, REN21). Renewable energy policies in a time of transition. Paris: IRENA, OECD/IEA, REN21; 2018. Available from: http://www.irena.org/-/media/Files/IRENA/Agency/Publication/2018/Apr/IRENA_IEA_REN21_Policies_2018.pdf

49. Kruger W, Eberhard A. Renewable energy auctions in sub-Saharan Africa: Review, lessons learned and recommendations. Cape Town: GSB, University of Cape Town; 2018. http://www.gsb.uct.ac.za/files/RenewableEnergyAuctionsSSA.pdf

50. ERC, CSIR, IFPRI. The developing energy landscape in South Africa: Technical report. Cape Town: Energy Research Centre, University of Cape Town; 2017. http://bit.ly/2ABamkz

51. Africa Progress Group Panel. Power people planet: Seizing Africa's energy and climate opportunities. Abeokuta, Nigeria: Africa Progress Group; 2015. Available from: https://www.seforall.org/sites/default/files/l/2015/06/APP_REPORT_2015_FINAL_low1.pdf

52. Bazilian M, Nussbaumer P, Rogner HH, Brew-Hammond A, Foster V, Pachauri S, et al. Energy access scenarios to 2030 for the power sector in sub-Saharan Africa. Util Policy. 2012;20:1–16. https://doi.org/10.1016/j.jup.2011.11.002

53. Levin K, Cashore B, Bernstein S, Auld G. Overcoming the tragedy of super wicked problems: Constraining our future selves to ameliorate global climate change. Policy Sci. 2012;45:123–152. https://doi.org/10.1007/s11077-012-9151-0

54. Monbiot G. Requiem for a crowded planet [webpage on the Internet]. c2009 [cited 2011 Oct 08]. Available from: http://www.monbiot.com/archives/2009/12/21/requiem-for-a-crowded-planet/

55. Pan J. Emissions rights and their transferability: Equity concerns over climate change mitigation. Int Environ Agreements Politics Law Econ. 2003;3:1–16. https://doi.org/10.1023/A:1021366620577

56. Mwandosya MJ. Survival emissions: A perspective from the South on global climate change negotiations. Dar es Salaam / Bangkok: Dar es Salaam University Press / Centre for Energy, Environment, Science and Technology; 2000.

57. Smith KR. The natural debt: North and South. In: Giambelluca T, Herderson-Sellers A, editors. Climate change: Developing southern hemisphere perspectives. New York: John Wiley; 1996. p. 423–448.

58. Ramaphosa C. Statement by His Excellency President Cyril Ramaphosa during the Open Session of the 10th BRICS Summit, 2018 July 26; Sandton International Convention Centre, Johannesburg, South Africa. Available from: http://www.dirco.gov.za/docs/speeches/2018/cram0726.html

59. Cox RW. Social forces, states and world orders: Beyond international relations theory. Millennium. 1981;10:126–155. https://doi.org/10.1177/03058298810100020501

60. Simon R, Hall S. Gramsci's political thought. London: Lawrence & Wishart; 2002.

61. Winkler H. Reducing energy poverty through carbon tax revenues in South Africa. J Energy South Africa. 2017;28:12–26. https://doi.org/10.17159/2413-3051/2017/v28i3a2332

62. Schwab K. The Fourth Industrial Revolution. London: Penguin; 2017.

63. Mason P. Post-capitalism: A guide to our future. London: Penguin; 2015.

64. Harari YN. Homo Deus: A brief history of tomorrow. London: Harvill Secker; 2016.

65. Srnicek N, Williams A. Inventing the future: Post-capitalism and a world without work. London: Verso Books; 2015.

66. Winkler H, Boyd A, Torres Gunfaus M, Raubenheimer S. Reconsidering development by reflecting on climate change. Int Environ Agreements. 2015;15:369–385. https://doi.org/10.1007/s10784-015-9304-7

67. Southern Africa Labour and Development Research Unit (SALDRU). Income comparison tool [webpage on the Internet]. No date [cited 2018 Oct 29]. Available from: http://www.saldru.uct.ac.za/income-comparison-tool/

Alpha and sigma taxonomy of *Pan* (chimpanzees) and Plio-Pleistocene hominin species

AUTHOR:
J. Francis Thackeray[1]

AFFILIATION:
[1]Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Francis Thackeray

EMAIL:
mrsples@global.co.za

KEYWORDS:
biological species constant; species variation; conspecificity; morphometrics; taxonomy

HOW TO CITE:
Thackeray JF. Alpha and sigma taxonomy of *Pan* (chimpanzees) and Plio-Pleistocene hominin species. S Afr J Sci. 2018;114(11/12), Art. #a0291, 2 pages. https://dx.doi.org/10.17159/sajs.2018/a0291

PUBLISHED:
27 November 2018

A fundamental question in biology, and more specifically in palaeontology, is 'how much variation is there within a biological species?' To answer that question, it is necessary to define a species, notably in a way that can be applied in palaeontological contexts. Recognising that boundaries between taxa may not always be clear, an appeal has been made for a probabilistic definition of a species[1-3], based on pairwise comparisons of specimens and morphometric analyses using least squares linear regression analysis associated with a general equation of the form $y=mc+c$, where x and y are linear dimensions of a skeletal element such as a cranium[4]. The degree of scatter around the regression equation (associated with morphology) is quantifiable using the log of the standard error of the m co-efficient (log sem). Here it is shown how this morphometric approach can be applied to cranial specimens attributed to two extant species of *Pan*, and to extinct Plio-Pleistocene hominins in a temporal sequence, indicating the lack of clear boundaries between species, thereby challenging the prevailing concept of alpha taxonomy[5] which assumes discrete entities. An appeal is made for an alternative concept, namely sigma taxonomy.[3]

Applications and a probabilistic definition of a species

The approach has been applied to measurements obtained from more than 70 taxa[1], and more recently to measurements of crania of *Pan troglodytes*, the common chimpanzee, and also to those of *P. paniscus*, the bonobo[2,6]. The results were remarkable in the sense that, in the case of both species analysed separately (using alpha taxonomy), a mean log sem value of -1.6 was obtained for conspecific pairs. The data confirmed a hypothesis proposed by Thackeray[1] that -1.61 for mean log sem values constitutes an approximation of a biological species constant (T), relating to a central tendency for the degree of variation within a species. An associated standard deviation for this proposed biological constant was given as 0.1 when using more than 2000 regression analyses for pairwise comparisons of specimens of the same species.[2]

A mean log sem value of -1.61±0.1 based on log sem statistics was considered to be a probabilistic definition of a species[2], relating to the degree of variability typically expressed within a single (extant) species.

Application to Plio-Pleistocene crania

The approach has been applied to cranial measurements of Plio-Pleistocene hominins.[4] Here, attention is restricted to five well-preserved and almost complete crania which have been attributed either to the genus *Australopithecus* or to the younger genus *Homo*. The five specimens and associated data are given in Table 1, in chronological order, from 1.6 million years ago (mya) to 2.5 mya.

Table 1: The sample of five almost complete Plio-Pleistocene cranial specimens

Specimen	Age	Taxon	Provenance
KNM-ER 3733	1.6 mya	*Homo erectus*	Turkana Basin, Kenya
OH 24	1.8 mya	*Homo habilis*	Olduvai, Tanzania
KNM-ER 1813	1.9 mya	*Homo habilis*	Turkana Basin, Kenya
Sts 5	2.1 mya	*Australopithecus africanus*	Sterkfontein, South Africa
Sts 71	2.5 mya	*Australopithecus africanus*	Sterkfontein, South Africa

A matrix of log sem values, based on pairwise comparisons of cranial measurements of these specimens[4] is given in Figure 1.

	Sts 71	Sts 5	ER 1813	OH 24	ER 3733	Key	LOG SEM SPECTRUM
Sts 71		-1.691	-1.546	-1.63	-1.526		-1.8
Sts 5	-1.643		-1.603	-1.506	-1.558		-1.7
ER 1813	-1.507	-1.627		-1.701	-1.844		-1.6
OH 24	-1.566	-1.509	-1.685		-1.723		-1.5
ER 3733	-1.318	-1.399	-1.656	-1.546			-1.4

Figure 1: Log sem values for pairwise comparisons of five Plio-Pleistocene hominin crania, dated between 1.6 and 2.5 million years ago. The mean log sem for the entire database is -1.59±0.12 which is almost identical to the values for conspecific comparisons of modern *Pan paniscus* (-1.61±0.1) and for *P. troglodytes* (-1.61±0.1).

The mean log sem value for the data set is -1.59±0.12 (n=20 pairwise comparisons). This result for a *temporal* sequence is remarkable in the sense that it expresses almost exactly the degree of variability that is found *spatially* in two modern species of chimpanzees (-1.61±0.1), examined more specifically below.

Four sets of independent data of mean log sem values, published by Gordon and Wood[6] and discussed by Thackeray and Dykes[2], are given in Table 2 for conspecific chimpanzees (male and female individuals are considered separately), to demonstrate consistency in the mean value of log sem for conspecific comparisons.

Table 2: A set of mean log sem values for pairwise comparisons of conspecific *Pan troglodytes* and *P. paniscus*

Log sem	Comparison
-1.61 ± 0.087	Female–female comparisons of *Pan paniscus*
-1.62 ± 0.095	Male–male comparisons of *Pan paniscus*
-1.62 ± 0.100	Female–female comparisons of *Pan troglodytes*
-1.60 ± 0.109	Male–male comparisons of *Pan troglodytes*
-1.61 ± 0.1	**Mean, n>2000 regression analyses**

Comparisons

On the basis of log sem statistics, it is evident that the spectrum of variability through *evolutionary time* (from 2.5 to 1.6 mya) in five Plio-Pleistocene hominins (mean log sem = -1.59±0.12) is comparable to the spectrum of variability in *geographical space* (-1.61±0.1) at the present time in *Pan paniscus* to the south of the Congo River. It is also comparable to the spectrum of variability in geographical space (-1.61±0.1) in *Pan troglodytes* to the north of that river.

Notably, the degree of variability (mean log sem = -1.61±0.1) in each of the two species of *Pan* developed within a period of (at least) one million years since the time of their divergence. However, when *P. troglodyes* and *P. paniscus* were compared with each other, log sem values did not show a clearly distinct separation.[2,6] This finding is consistent with genetic evidence for hybridisation between *P. troglodytes* and *P. paniscus* within the last million years.[7]

Hypotheses and a definition

Using the results presented here for two species of chimpanzees which diverged about 1 million years ago, and also for five Plio-Pleistocene hominins in a sequence within about one million years, the following hypotheses are presented:

H1: There is no clear boundary between *P. troglodytes* and *P. paniscus.*

H2: There is no clear boundary between certain species attributed to *Australopithecus* and *Homo*.

H3: Certain hominin species attributed to *Australopithecus* and to *Homo* were capable of interbreeding within a period of a million years (a spectrum of time between 1.6 and 2.5 million years ago).

These observations and hypotheses serve to underscore the importance of developing a probabilistic definition of a species that relates to sigma taxonomy, where sigma is the Greek letter for S (Σ) standing for the concept of a spectrum[3,8,9], as opposed to alpha taxonomy which assumes clear boundaries between species[5]. A formal definition for sigma taxonomy is: 'The classification of taxa in terms of probabilities of conspecificity, without assuming distinct boundaries between species'.

Acknowledgements

This work is supported by the National Research Foundation of South Africa and the DST/NRF Centre of Excellence for the Palaeosciences.

References

1. Thackeray JF. Approximation of a biological species constant? S Afr J Sci. 2007;103:489.
2. Thackeray JF, Dykes S. Morphometric analyses of hominoid crania, probabilities of conspecificity and an approximation of a biological species constant. Homo. 2016;67(1):1–10. http://dx.doi.org/10.1016/j.jchb.2015.09.003
3. Thackeray JF, Schrein CM. A probabilistic definition of a species, fuzzy boundaries and 'sigma taxonomy'. S Afr J Sci. 2017;113(5/6), Art. #a0206, 2 pages. http://dx.doi.org/10.17159/sajs.2017/a0206
4. Thackeray JF, Odes E. Morphometric analysis of early Pleistocene African hominin crania in the context of a statistical (probabilistic) definition of a species. Antiquity. 2013;87(335):1–3. Available from: http://antiquity.ac.uk/projgall/thackeray335/
5. Mayr E, Linsley EG, Usinger RL. Methods and principles of systematic zoology. New York: McGraw-Hill; 1953.
6. Gordon AD, Wood BA. Evaluating the use of pairwise dissimilarity metrics in paleoanthropology. J Hum Evol. 2013;65:465–477. https://doi.org/10.1016/j.jhevol.2013.08.002
7. De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354(6311):477–481. https://doi.org/10.1126/science.aag2602
8. Thackeray JF. Sigma taxonomy in relation to palaeoanthropology and the lack of clear boundaries between species. Proc Eur Soc Stud Hum Evol. 2015;4:220.
9. Thackeray JF. *Homo habilis* and *Australopithecus africanus*, in the context of a chronospecies and climatic change. In: Runge J, editor. Changing climates, ecosystems and environments within arid southern Africa and adjoining regions: Palaeoecology of Africa 33. Leiden: CRC Press/Balkema; 2015. p. 53–58.

Are managed pollinators ultimately linked to the pollination ecosystem service paradigm?

AUTHOR:
Ruan Veldtman[1,2]

AFFILIATIONS:
[1]South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa
[2]Department of Conservation Ecology and Entomology, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Ruan Veldtman

EMAIL:
r.veldtman@sanbi.org.za

HOW TO CITE:
Veldtman R. Are managed pollinators ultimately linked to the pollination ecosystem service paradigm? S Afr J Sci. 2018;114(11/12), Art. #a0292, 4 pages. https://dx.doi.org/10.17159/sajs.2018/a0292

KEYWORDS:
Apis mellifera; forage provision; agricultural input

PUBLISHED:
27 November 2018

Crop pollination performed by wild pollinators is arguably the best understood animal-based ecosystem service. Pollination by wild pollinators originating from natural habitats is recognised as an important ecosystem service; in contrast, managed pollinators – overwhelmingly represented by *Apis mellifera* (the European honey bee) – are regarded by most as an agricultural input.[1-3] Globally, both wild and managed insect pollinators are important for crops requiring pollination.[2,4-6] The principal difference between these two pollination services is that wild pollinators are residents while managed pollinators are imported for crop pollination (Figure 1). However, there are cases in which managed honey bee hives are kept at permanent locations. These managed honey bees, akin to resident wild pollinators, follow the available forage resources found within flying distance.

Globally, the demand for insect pollinated crops is increasing at a phenomenal rate and as the human population continues to increase and improve its standard of living, demand can only increase.[1] Global food demand has in the past been met by monoculture environments transformed during the green revolution[7], resulting in a predominance of this agroecosystem at the expense of more diverse crop-natural margins, at least in developed countries.[4] Consequently, before the focus on wild pollinators, research on managed honey bees dominated the crop pollination literature.[4,6]

However, in the last two decades, there has been a complete turnaround. Most published studies on insect crop pollination services are introduced by stating three 'near universal' facts: (1) animal pollinators provide an important ecosystem service but these pollinators are now threatened, (2) managed honey bees are commonly used to mediate their loss but are themselves declining and (3) more effort must be placed in making crop ecosystems more pollinator friendly for wild pollinators to restore the free service they provide. There are now numerous studies and global meta-analyses that document a concurrent increase in pollinator diversity and crop yields as a result of better pollination (and thus improved food security) with practices that are ecologically friendly and promote on-farm wild pollinator conservation.[2,4,8] These reports have led to calls to reverse native pollinator declines by improving the on-farm environment for pollinators.[4,5,8]

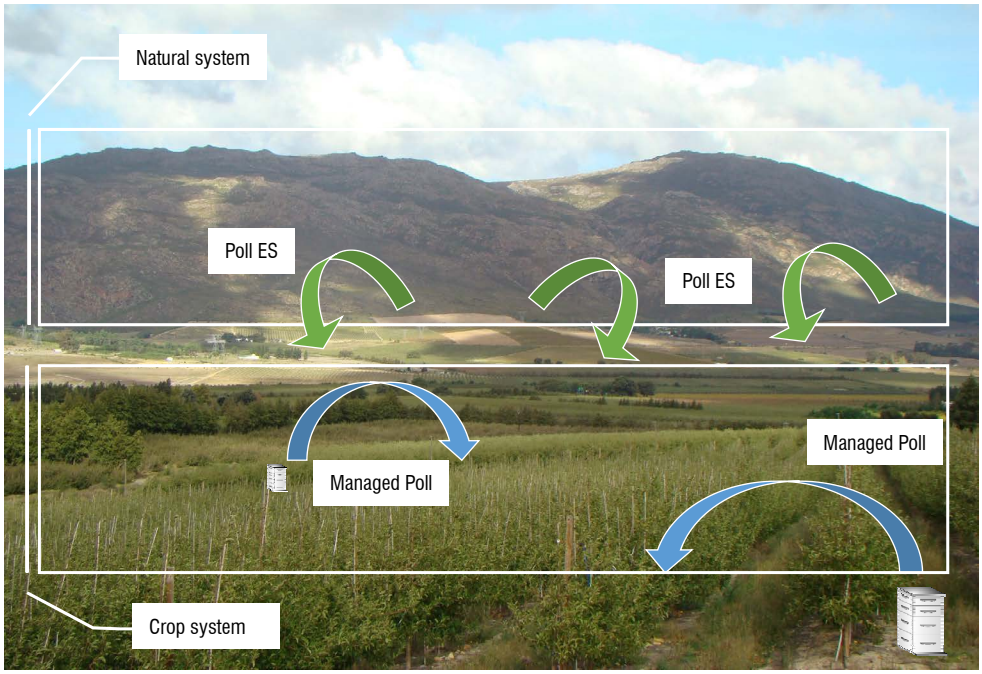


Figure 1: A comparison of pollination services – wild and managed pollination services offered to a deciduous crop farmer. A grower's options for using pollination ecosystem services is strongly determined by landscape context and field size. In contrast, rented honey bees do not have this constraint but are a paid-for service.

While the current research effort into pollinator conservation, ecologically intensifying agriculture and resulting initiatives are commendable, it does not change the fact that there is a forced dependence on managed pollinators.[3,9,10] For many global crop hectares, there is no access to viable wild pollination services[6,11] and restoration efforts will not be able to replace the contribution of managed pollinators[12,13]. I argue that the implications of the necessary contribution of managed pollinators is not fully comprehended by many.

Melathopoulos et al.[3] and others consider managed pollinators to more closely resemble an agricultural input, because in most parts of the world these managed pollinators are often non-native[4], have only a temporary dependence on the habitat surrounding the fields they pollinate[3], and instead are highly dependent on manufactured food substances (such as sucrose and processed plant proteins) and chemical inputs such as miticides and

antibiotics[14,15]. It is precisely this strong human dependence that allows these species to function in highly intensified agro-ecological landscapes that would otherwise not support comparable levels of pollination ecosystem services.[6,16] For managed pollinators that are in the front lines of intensive agriculture, are human interventions such as disease control[17] and optimising dietary demand[18] enough to ensure a sustainable pollination service?

There is widespread consensus that managed honey bees also benefit from a varied landscape with not only flower strips, but also the proximity of natural vegetation. Several studies have shown that the availability of pollinator habitat can be used as a gauge for pollination ecosystem services[8,11] and some of these studies explicitly show that bee habitat is not only a year-round habitat for wild bees, but also a food resource for managed honey bees[9,13]. However, the issue discussed here is not the provision of forage for managed honey bees at a single site, but rather, what other sites are needed to support these managed pollinators throughout a year?

A recent review on pollination ecosystem services in South Africa[19] concluded that more research is needed to document wild pollination services, citing the under-reporting of wild versus managed honey bees pollinating specific crops as the biggest stumbling block. Nonetheless, it is well documented that managed honey bees play a pivotal role in the pollination of South Africa's crops.[19,20] Given the case studies already presented which show a predominance of honey bee flower visitation (cited in Melin et al.[19]), and because most commercial agricultural crops are intensively grown without diversified flower strips or proximity to natural vegetation to support wild honey bee colonies[11,21,22], we can assume that there is no wild pollinator replacement for the managed pollination services currently offered in South Africa[22-25]. Thus, irrespective of international trends, South Africa's biggest threat to meeting the ever-increasing crop pollination demand seems to be an insufficient number of managed honey bees for hive rental services.

It is my opinion that current international literature overemphasises the importance of pollination ecosystem services because managed pollinators are simply considered an agricultural input, and their dependence on off-farm sites containing natural and semi-natural flower resources (required as forage when they are not pollinating crops) is not being considered. Such forage can be seen as a provisioning ecosystem service[26], similar to providing forage for free-range domestic livestock or game farming. In the absence of forage sources, beekeepers rely on sub-optimal and expensive artificial feed.[24,27] While it is true that managed pollinators are owned and transported to crops by a beekeeper (pollinator manager) – which cannot be considered a pollination ecosystem service – the forage required to sustain these colonies during a year *does* require a forage provisioning service. Usually a range of foraging areas (typically not owned by the beekeeper) are used, thereby allowing beekeepers to track floral pulses and allow high-density beekeeping.

Some countries, such as China and Argentina, are rich in forage and typically dominate other countries' honey sales. South Africa is a forage poor country because most of the country experiences low rainfall and the majority of honey is now imported from these forage rich countries.[22] There are even further imbalances in areas such as the Western Cape where there is a very high crop pollination demand, such that forage is usually not used for honey production but for supporting managed pollination services (hives are rented for pollination). For this reason, the dependence of managed pollinators on the natural and human modified environment to provide forage in the form of pollen and nectar, should be thought of as a provisioning ecosystem service[26], or, more specifically, as the ecosystem service of off-farm (i.e. when not rented for on-farm pollination) forage provision for managed pollinators (Figure 2).

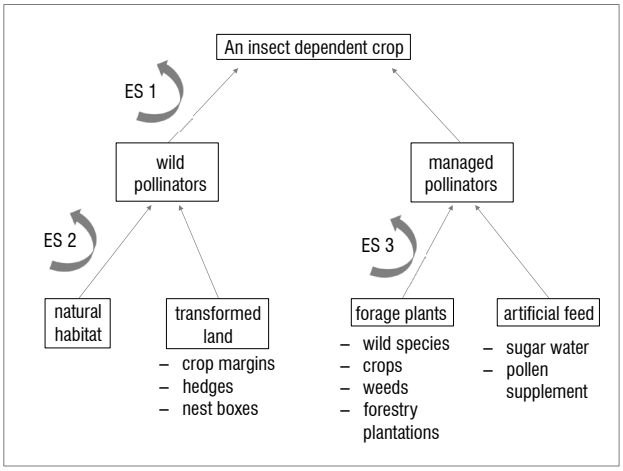


Figure 2: Framework to illustrate pollination of insect-dependent crops. Pollination by wild pollinators is an ecosystem service (ES 1), while managed pollinators placed near the crop via human intervention is not. Wild pollinators are supported by natural habitat which provides forage and nesting sites, i.e. the ecosystem service of biodiversity maintenance (ES 2), or can be accommodated in transformed land which has been ecologically enhanced by keeping natural, exotic or crop species nearby. In comparison, the food requirements of managed pollinators can be met by keeping them at sites with plant species (natural, exotic or crops) that provide nectar and/or pollen, i.e. a forage provisioning ecosystem service (ES 3), or they can be fed artificial feed as a forage replacement.

In this case, even human-dominated ecosystems that contain foraging plants can support managed pollination (Figure 3), with both the beekeeper, agricultural producers, and society at large benefiting from the maintenance and availability of such forage resources. The ecosystem service of forage provision can thus be seen to subsidise the rental cost of the managed honey bee pollination service.[27] Consequently, similar to classical ecosystem service use, if the private landowner is not incentivised to maintain forage resources, the beekeeper and everyone else who benefits, will lose a resource, resulting in knock-on effects along the supply chain to the consumer of pollinator-dependent products. For South Africa which is a major fruit exporter, this loss of resources would also mean a loss of export revenue. If it is only a matter of user-pays, beekeepers would have to pay for forage used, thus limiting the number of colonies kept and/or increasing the rental price for pollination. Growers in turn would experience higher costs or production deficits and would at best sell the same quantity of produce at a higher price, resulting in a loss of consumer welfare.[28] De Lange et al.[27] show the cost of replacing forage by either restoring natural vegetation or providing artificial feed is orders of magnitude greater than the cost of renting these managed pollination services. Therefore, the cost of managed pollination rental is a gross underestimate of the value of managed crop pollination services to food production.[20,27,29] It can thus be seen that off-farm forage for managed pollinators is also a critical resource to support sustainable crop pollination and food security.

This view, however, will not be readily accepted internationally by scientists working on pollination ecosystem services because in most places *Apis mellifera* is not native and/or is heavily dependent on beekeepers for forage supplements.[3] Furthermore, the presence of managed pollinators (including *Bombus terrestris*) can have negative implications for the viability of wild congeners.[30,31] Nonetheless, there is a strong case to be made where the honey bee is native and receives minimum forage supplements – as is the case in South Africa. In fact, one could argue that because European countries (where *Apis mellifera* is also native) have not in the past explicitly considered the service of forage provision for managed honey bees when their agricultural systems were becoming very intensive (and uninhabitable for wild pollinators), there was no planning for forage areas and as a consequence they now have a very

dependent managed pollinator species. In contrast, South Africa makes use of 'robust' beekeeping with minimum input from beekeepers and there is an arbitrary separation between wild and managed honey bee colonies.[22]

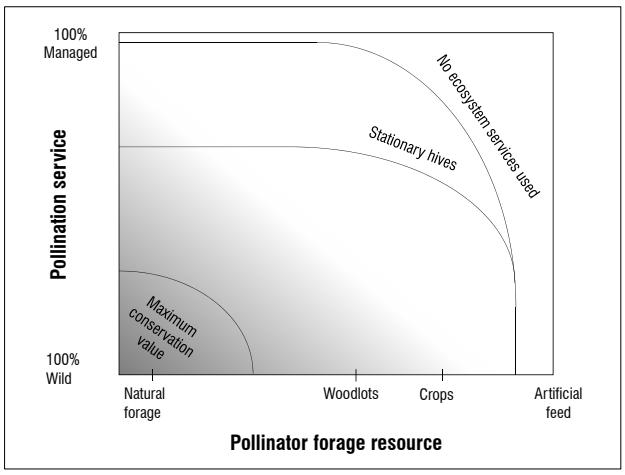


Figure 3: Hypothetical relationship between pollination services, floral resource use and the degree of harnessing ecosystem services for crop pollination (degree of grey fill indicates level of ecosystem service use). There is not a binary ecosystem service dependence but rather a gradient of pollination services and forage resources used, that at times may even have conservation importance (i.e. native managed pollinator and native plant species, respectively). Managed honey bee hives kept at a permanent forage site that happens to be near a pollinator dependent crop is an interesting case in which a hybrid crop pollination service is provided.

Mitigation measures to secure wild pollinators are now globally well established – both scientifically and through policy.[5,32,33] South Africa, however, cannot simply follow international trends by over-promoting wild pollination services at the expense of resources for managed honey bees. Thus, the resources used when honey bees are not being rented for crop pollination also need to be accounted for[25,26,34,35] and it is here where there is a lack of comprehensive information for South Africa; globally, there is no consideration of how many managed honey bee hives are required annually for the pollination of all pollinator-dependent crops, not to mention how these hives are supported.

On-farm forage resources for wild pollinators is certainly important, but it is equally certain that off-farm forage resources for migratory-managed pollinators are at least as important. Regarding managed pollinators simply as an ecologically inert agricultural input discounts the ecosystem resources on which they rely, which then weakens policy initiatives aimed at improving crop pollination and food security. Consider for a moment how the recent drought in the Western Cape has shown the importance of maintaining catchments free of alien invasive plants to maximise water recharge (a classic ecosystem service).[36] However, in the same region where there is very limited managed honey bee forage provisioning, while eucalyptus stands are being cleared to liberate water resources, the forage provisioning service these plants provide is at the same time being destroyed.[5] Careful consideration must be given to the trade-off between the ecosystem services of water provision and that of forage provision.[27] I propose that explicitly considering forage provision for managed pollinators as an ecosystem service (e.g. Mensah et al.[26] and Melin et al.[35]), will help correct the perception that sustainable managed pollination is only an agricultural issue.

Acknowledgements

I thank Mike Allsopp for several years of discussion on the content, as well as Colleen Seymour, Breno Freitas, Bernard Vaissière, Tlou Masehela, Michael Samways and an anonymous reviewer for comments on previous drafts. I declare no conflict of interest in publishing this work.

References

1. Aizen MA, Harder LD. The global stock of domesticated honey bees is growing slower than agricultural demand for pollination. Curr Biol. 2009;19:1–4. https://doi.org/10.1016/j.cub.2009.03.071
2. Garibaldi LA, Steffan-Dewenter I, Winfree R, Aizen M, Bommarco R, Cunningham SA, et al. Wild pollinators enhance fruits set of crops regardless of honey bee abundance. Science. 2013;339(6127):1608–1611. http://dx.doi.org/10.1126/science.1230200
3. Melathopoulos AP, Cutler GC, Tyedmers P. Where is the value in valuing pollination ecosystem services to agriculture? Ecol Econ. 2015;109:59–70. https://doi.org/10.1016/j.ecolecon.2014.11.007
4. Klein A, Vaissière BE, Cane JH, Steffan-Dewenter I, Cunningham SA, Kremen C, et al. Importance of pollinators in changing landscapes for world crops. Proc Biol Sci. 2007;274(1608):303–313. http://dx.doi.org/10.1098/rspb.2006.3721
5. Potts SG, Imperatriz-Fonseca VL, Ngo HT, Biesmeijer JC, Breeze TD, Dicks LV, et al., editors. Summary for policymakers of the assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production. Bonn: Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services; 2016. p. 1–36.
6. Cunningham SA, Fournier A, Neave MJ, Le Feuvre D. Improving spatial arrangement of honeybee colonies to avoid pollination shortfall and depressed fruit set. J Appl Ecol. 2016;53:350–359. https://doi.org/10.1111/1365-2664.12573
7. Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S. Agricultural sustainability and intensive production practices. Nature. 2002;418(6898):671–677. https://doi.org/10.1038/nature01014
8. Garibaldi LA, Carvalheiro LG, Vaissière BE, Gemmill-Herren B, Hipólito J, Freitas BM, et al. Mutually beneficial pollinator diversity and crop yield outcomes in small and large farms. Science. 2016;351(6271):388–391. https://doi.org/10.1126/science.aac7287
9. Aebi A, Vaissière BE, Van Engelsdorp D, Delaplane KS, Roubik DW, Neumann P. Back to the future: *Apis* versus non-*Apis* pollination. Trends Ecol Evol. 2012;27(3):142–143. http://dx.doi.org/10.1016/j.tree.2011.11.017
10. Calderone NW. Insect pollinated crops, insect pollinators and US agriculture: Trend analysis of aggregate data for the period 1992–2009. PLoS ONE. 2012;7(5), e37235, 27 pages. https://doi.org/10.1371/journal.pone.0037235
11. Carvalheiro LG, Seymour CL, Veldtman R, Nicolson SW. Pollination services decline with distance from natural habitat even in biodiversity-rich areas. J Appl Ecol. 2010;47(4):810–820. http://dx.doi.org/10.1111/j.1365-2664.2010.01829.x
12. Kremen C, Daily GC, Klein A-M, Scofield D. Inadequate assessment of the ecosystem service rationale for conservation: Reply to Ghazoul. Conserv Biol. 2008;22(3):795–798. https://doi.org/10.1111/j.1523-1739.2008.00940.x
13. Klein A-M, Brittian C, Hendrix SD, Thorp R, Williams N, Kremen C. Wild pollination services to California almond rely on semi-natural habitat. J Appl Ecol. 2012;49:723–732. https://doi.org/10.1111/j.1365-2664.2012.02144.x
14. Southwick EE, Pimentel D. Energy efficiency of honey production by bees. Bioscience. 1981;31:730–732. https://doi.org/10.2307/1308779
15. Kendall A, Yuan J, Brodt SB. Carbon footprint and air emissions inventories for US honey production: Case studies. Int J Life Cycle Assess. 2013;18:392–400. https://doi.org/10.1007/s11367-012-0487-7
16. Ghazoul J. Challenges to the uptake of the ecosystem service rationale for conservation. Biol Conserv. 2007;21(6):1651–1652. https://doi.org/10.1111/j.1523-1739.2007.00758.x
17. Van Engelsdorp D, Meixner MD. A historic review of managed honey bee populations in Europe and the United States and the factors that may affect them. J Invert Pathol. 2010;103:580–595.
18. Alaux C, Ducloz F, Crauser D, Le Conte Y. Diet effects on honeybee immunocompetence. Biol Lett. 2010;6(4):562–565. http://dx.doi.org/10.1098/rsbl.2009.0986
19. Melin A, Rouget M, Midgley J, Donaldson JS. Pollination ecosystem services in South African agricultural systems. S Afr J Sci. 2014;110(11/12), Art. #2014-0078, 9 pages. http://dx.doi.org/10.1590/sajs.2014/20140078

20. Allsopp MH, De Lange WJ, Veldtman R. Valuing insect pollination services with cost of replacement. PLoS ONE. 2008;3(9), e3128, 8 pages. http://dx.doi.org/10.1371/journal.pone.0003128

21. Carvalheiro LG, Veldtman R, Shenkute AG, Tesfay GB, Walter C, Pirk W, et al. Natural and within-farmland biodiversity enhances crop productivity. Ecol Lett. 2011;14(3):251–259. http://dx.doi.org/10.1111/j.1461-0248.2010.01579.x

22. Allsopp MH, Veldtman R. Managed honeybee industry and honeybee forage dependencies: The cost of being important. S Afr Bee J. 2012;84(3):160–165.

23. Brand M. Pollination ecosystem services to hybrid onion seed crops in South Africa [PhD thesis]. Stellenbosch: Stellenbosch University; 2014.

24. Hutton-Squire JP. Historical and current relationship between the honeybee (*Apis mellifera*) and its forage in South Africa [MSc thesis]. Stellenbosch: Stellenbosch University; 2015.

25. Masehela TS. An assessment of different beekeeping practices in South Africa based on their needs (bee forage use), services (pollination services) and threats (hive theft and vandalism) [PhD thesis]. Stellenbosch: Stellenbosch University; 2017.

26. Mensah S, Veldtman R, Seifert T. Potential supply of floral resources to managed honey bees in natural mistbelt forests. J Environ Manage. 2017;189:60–167. https://doi.org/10.1016/j.jenvman.2016.12.033

27. De Lange WJ, Veldtman R, Allsopp MH. Valuation of pollinator forage services provided by *Eucalyptus cladocalyx*. J Environ Manage. 2013;125:12–18. http://dx.doi.org/10.1016/j.jenvman.2013.03.027

28. Gallai N, Salles J, Settele J, Vaissière B. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecol Econ. 2009;68(3):810–821. http://dx.doi.org/10.1016/j.ecolecon.2008.06.014

29. Winfree R, Gross BJ, Kremen C. Valuing pollination services to agriculture. Ecol Econ. 2011;71:80–88. http://dx.doi.org/10.1016/j.ecolecon.2011.08.001

30. Fürst M, McMahon D, Osborne J, Paxton R, Brown M. Disease associations between honeybees and bumblebees as a threat to wild pollinators. Nature. 2014;506:364–366. https://doi.org/10.1038/nature12977

31. Goulson D. Effects of introduced bees on native ecosystems. Annu Rev Ecol Evol Syst. 2003;34(1):1–26. http://dx.doi.org/10.1146/annurev.ecolsys.34.011802.132355

32. Brown MJF, Dicks LV, Paxton RJ, Baldock KCR, Barron AB, Chauzat M, et al. A horizon scan of future threats and opportunities for pollinators and pollination. PeerJ. 2016;4, e2249, 20 pages. https://doi.org/10.7717/peerj.2249

33. Potts SG, Imperatriz-Fonseca V, Ngo HT, Aizen MA, Biesmeijer JC, Breeze TD, et al. Safeguarding pollinators and their values to human well-being. Nature. 2016;540(7632):220–229. http://dx.doi.org/10.1038/nature20588

34. Houdet J, Veldtman R. How to account for bees and pollinators. In: Atkins J, Atkins B, editors. The business of bees: An integrated approach to bee decline and corporate responsibility. Sheffield: Greenleaf Publishing; 2016. p. 188–197.

35. Melin A, Rouget M, Colville JF, Midgley J, Donaldson JS. Assessing the role of dispersed floral resources for managed bees in providing supporting ecosystem services for crop pollination. Peer J. 2018;6, e5654, 23 pages. https://doi.org/10.7717/peerj.5654

36. Dzikiti S, Gush MB, Le Maitre DC, Maherry A, Jovanovic NZ, Ramoelo A, et al. Quantifying potential water savings from clearing invasive alien *Eucalyptus camaldulensis* using in situ and high resolution remote sensing data in the Berg River Catchment, Western Cape, South Africa. Forest Ecol Manage. 2016;361:69–80. https://doi.org/10.1016/j.foreco.2015.11.009

Research biobanks: A two-faced future

AUTHORS:
Marco Capocasa[1]
Valentina Dominici[2]
Fabrizio Rufo[1,2]

AFFILIATIONS:
[1]Italian Institute of Anthropology, Rome, Italy
[2]Department of Environmental Biology, Sapienza University of Rome, Rome, Italy

CORRESPONDENCE TO:
Marco Capocasa

EMAIL:
marco.capocasa@yahoo.it

KEYWORDS:
biobanking; accessibility; biological samples; data; network

HOW TO CITE:
Capocasa M, Dominici V, Rufo F. Research biobanks: A two-faced future. S Afr J Sci. 2018;114(11/12), Art. #5402, 3 pages. https://dx.doi.org/10.17159/sajs.2018/5402

PUBLISHED:
27 November 2018

The possibility of preserving human tissue separated from the body, from which to extract clinical information, even on large numbers of individuals with similar clinical conditions, represents a great opportunity for the progress of biomedicine. To date, it is virtually impossible to hypothesise all the future uses of such tissue. However, the decisive role of these biological materials in the understanding and resolution of questions regarding the origin and development of certain genetic diseases is well recognised.

The progress of genetic and biotechnological research has led to the proliferation of collections of biological materials by institutions called 'biobanks': repositories that store human biological samples, with or without linking them to genetic or clinical data.[1] Usually biobanks are part of large public research centres, small hospitals and pharmaceutical companies. They conduct their activities following different, and not always standardised, storage and conservation protocols. Because of this enormous heterogeneity of structures, materials and methodologies, research biobanks have been subjected to profound ambiguity and fragmentation, as well as uncertainty in terms of regulation. This situation has encouraged discussion on the legal rules regarding their activities, particularly for the protection of donors' rights.[2-4]

In the 21st century, the technological progress in automation and archival informatisation and the development of the World Wide Web boosted a radical revolution of biobanking. Marked changes have occurred in the context of the management of biological samples, particularly in regard to their transport and storage. However, progress has also been made in regard to the ethical and legal requirements necessary to ensure the privacy and safety of donors and the long-term availability of these materials. Collecting samples in a more accurate and ethical way may be a crucial contribution in the advancement of clinical and biomedical trials and, more generally, in the creation of more informative data sets. Such progress is also associated with a substantial increase in the production of data as a consequence of the collection of pathological, epidemiological, environmental and ethno-social information of the donors.

The collection of biological materials in compliance with ethical and legal standards could be seen as the main innovation of research biobanks. Firstly, collections of samples are virtually unusable today because of the lack of detail in the informed consent or because they were catalogued following inappropriate methods.[5] Overcoming these issues represents one of the main enhancements for the development of modern and more structured biobanks.

Biobanking: Between politics and economics

Research biobanks often operate in very different contexts and, in many cases, their activities have become of primary interest to government agencies. In fact, the relationship between biobanks and politics is strengthening as many governments are increasingly interested in the efficiency and standardisation of ethical-legal frameworks for the sampling of biological materials. There is reciprocity between scientists and politicians: the collection of more, and extensively useable, human biological samples helps the former to conduct more accurate studies and guarantees the support of the latter regarding their political choices on scientific research.

Biobanking has also revealed unprecedented business opportunities. In 2011, Global Industry Analysts released a global report on the biobank market, in which they predicted that the market in 2017 for high-quality human biological samples would be USD22.3 billion. Thus, the political and economic interests of biobanking are clear.

Scientific interests are obviously the first priority. There is no doubt that, globally, the development and growth of biobanks has exponentially increased opportunities for analysing and studying the collection of human biological samples and data extrapolated therefrom. This growth corresponds to an increase in research perspectives compared with those possible if the sampling and management of biological materials was still exclusively tied to the initiatives of individual research groups.[1] In addition to these new opportunities, the rapid evolution of biobanking has also created new challenges and obstacles. Particularly, researchers have difficulty in accessing biobank resources. The propensity of these institutions to share their samples and data with the scientific community is a controversial subject which encompasses a double challenge: for researchers, gaining access to biological samples and data, and, for biobanks, finding a balance between the scientific interests of researchers and donors' expectations. A first step towards a solution capable of satisfying both sides could be represented by a better and more flexible use of current forms of informed consent.[6,7] However, the requirement of a consent form including a section ensuring the sustainability of a wide accessibility to samples and data does not solve the problem related to economic interests, which are often hidden and protected by the scientific aim itself.

Although biobanks have suddenly become more 'open', several bioethical issues related to the sharing procedures of biological resources have emerged. Starting from the drafting of informed consent, biobanks must take into account a number of 'twists' in the definition of the section concerning the handling of samples and data. They should meet the requirements of ethical committees. They should also clearly state the hypothetical future uses of their resources. Moreover, they should enable potential donors to make truly informed decisions.

Accessibility to biobank resources is generally conditional on the fulfilment of specific, often very compelling, requirements and seems to be related to three main aspects.[8] The first aspect is the transparency between biobanks and applicants. Particularly, before allowing others to use their own materials, biobanks want to know the scientific aims of applicants. This request is strictly linked with specific sharing statements reported in the original consent form. It also provides a certain level of control by the biobanks regarding the scientific reliability and reputation

of applicants and their research groups, in order to reduce the risk of misuse of biobank resources. This is an ethical and technical approach to the management of scientific resources that can foster public trust in the work of these institutions, thus increasing willingness to participate in their activities.

Secondly, accessibility is related to the availability and origin of research funds. Biobanks seem to be more prone to collaborate with research groups that are publicly funded. This trend reflects the Organisation for Economic Co-operation and Development (OECD)'s recommendation in 2007 for open access to scientific resources, in which these resources are defined as 'public goods'. The OECD considers the sharing of these resources to be a means of enhancing public investment in scientific research. The availability of funds is a criterion adopted by biobanks in deciding whether or not to provide their resources to third parties. The presence of clauses directly linked to certain economic benefits for biobanks reveals their possible 'second nature' as institutions which also make a profit in providing bio-collection and storage services. However, it is not clear if this commercial nature operates as a sharing barrier, thereby lowering the risk of exploitation of their resources. As suggested by Caulfield and colleagues[9], sample and data sharing is a practice that can be influenced, or even hindered, by the introduction of private funding and by collaboration with private groups. In fact, the latter could have economic expectations concerning the use of resources produced with their money. Consequently, these groups could have entered into agreements governing such collaborations by acting as sharing barriers.

Thirdly, accessibility is related to co-authorship. Some biobanks require recognition as co-authors on publications resulting from research based on the analysis of their samples and data. Other authors have highlighted this typical bad practice in the sharing behaviour of research groups.[10-12] Clearly, it contributes to the spread of a climate of mistrust and a lower propensity for cooperation within the scientific community.

These findings suggest that both economic and academic aspects are involved in determining ways to manage the exploitation of biological resources by biobanks currently operating on a global scale. However, biobanks can differ completely in goals and outcomes and diametrically opposed visions can coexist within the same biobank. We should also take into account that biobanks are regulated on the basis of both national and international rules (such as the Transatlantic Trade and Investment Partnership bilateral agreement between the USA and the European Union), which sometimes hinder harmonisation, with consequences that may affect national health services and, more generally, biomedical research.

Looking to tomorrow

The availability of high-quality samples, accompanied by detailed metadata, will be the decisive push for the discovery of previously unknown biomarkers, thus facilitating the definition of new and innovative therapies.[13,14] The only way to increase the number and variety of human biological samples is to increase the 'source of tissues' and the efficiency of their provision.[5] The future of biobanks will depend, first and foremost, on their ability to respond to this increased demand. However, biobanks will only be able to overcome this challenge if they can also significantly reduce sampling and distribution costs. If we look at biobanking from this point of view, its tomorrow will be a matter of business in which only those institutions that will be able to operate efficiently and sustainably will remain and compete.

In this scenario, biobank networks can be seen as the most promising strategy to try to facilitate accessibility to samples and their findability. In the USA, the Cooperative Human Tissue Network (CHTN) is a 'generalised biobank' capable of collecting any kind of human samples for any biomedical research. In Europe, the Biobanking and BioMolecular Resources Research Infrastructure (BBMRI-ERIC) has been in operation since 2013 and is probably the world's largest biobank network. BBMRI-ERIC provides access to human biological samples that are considered raw materials needed for the advancement and development of biomedicine in Europe. Consortia such as CHTN and BBMRI-ERIC are increasingly essential, particularly for those research groups that do not have enough funding and human capacity to maintain an efficient biobank. However, while these networks are organised and equipped with impressive numbers of high-quality samples, obviously they will not actually meet all the needs of research groups. This is because certain studies require a specific type of biological material and a large number of samples collected from a specific population. Thus, there will still be room for local and small biobanks.

In the near future, another aspect that will affect biobanks' choices in collecting samples, particularly regarding the type of tissue and the sample size, will be the so-called 'post-genomic revolution'. It will force researchers around the world towards increasingly specific sample requirements that will reverberate in the activity of biobanks, particularly in the need to develop more coordinated collections to satisfy specific scientific issues. De Souza and Greenspan[15] point out that this process has already begun, as evidenced by the advent of population biobanks, as well as biobanks which collect only DNA or focus their activity on a single pathology.

We can conclude that the future of biobanking is double-edged. Bright and, at the same time, full of obstacles. First of all, its evolution will depend on addressing some of the unsolved problems preventing the full exploitation of resources, mostly in regard to the lack of shared standards for collecting, cataloguing and managing samples. Another issue is the long-term sustainability of these institutions. The scientific community needs efficient biobanks to support long-term studies. However, this aspect would require a separate study of the critical points linked with the maintenance of such institutions and the efficiency of their activities in supporting actively and dynamically the research enterprise. Biobanks will also have to face the question of open access to their resources. As discussed above, there are still barriers to the sharing of samples and data between biobanks and researchers. The fact that the scientific community is fully aware of the importance of sharing does not necessarily mean that these barriers are harmful and unfair and that they should not exist. In fact, some of them play an important role, because they exist to guarantee the fundamental rights of donors and prevent the misuse of biological materials. Thus, respect for donors' rights should always be given consideration in trying to overcome these barriers, not as a means to unify all the procedures for accessing biological resources but to find a definition of a globally recognised operating standard.

Acknowledgements

This work was supported by the Istituto Italiano di Antropologia (www.isita-org.com).

References

1. Haga SB, Beskow LM. Ethical, legal, and social implications of biobanks for genetics research. Adv Genet. 2008;60:505–544. https://doi.org/10.1016/S0065-2660(07)00418-X
2. Novelli G, Pietrangeli I. I campioni biologici [Biological samples]. In: Rodotà S, Tallacchini M, editors. Trattato di biodiritto [Treatise of biolaw]. Milan: Giuffrè; 2010. p. 1027–1061. Italian.
3. Hoeyer KL. Size matters: The ethical, legal, and social issues surrounding large-scale genetic biobank initiatives. Nor Epidemiol. 2012;21:211–220. https://doi.org/10.5324/nje.v21i2.1496
4. Macilotti M. Informed consent and research biobanks: A challenge in three dimensions. In: Pascuzzi G, Izzo U, Macilotti M, editors. Comparative issues in the governance of research biobanks. Berlin: Springer; 2013. p. 143–161. https://doi.org/10.1007/978-3-642-33116-9_9
5. Somiari SB, Somiari RI. The future of biobanking: A conceptual look at how biobanks can respond to the growing human biospecimen needs of researchers. Adv Exp Med Biol. 2015;864:11–27. https://doi.org/10.1007/978-3-319-20579-3_2
6. Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet. 2012;13:415–431. https://doi.org/10.1146/annurev-genom-082410-101454
7. D'Abramo F. Biobank research, informed consent and society. Towards a new alliance? J Epidemiol Community Health. 2015;69:1125–1128. https://doi.org/10.1136/jech-2014-205215

8. Capocasa M, Anagnostou P, D'Abramo F, Matteucci G, Dominici V, Destro Bisol G, et al. Samples and data accessibility in research biobanks: An explorative survey. PeerJ. 2016;4, e1613, 18 pages. https://doi.org/10.7717/peerj.1613

9. Caulfield T, Burningham S, Joly Y, Master Z, Shabani M, Borry P, et al. A review of the key issues associated with the commercialization of biobanks. J Law Biosci. 2014;1:94–110. https://doi.org/10.1093/jlb/lst004

10. Vogeli C, Yucel R, Bendavid E, Jones LM, Anderson MS, Louis KS, et al. Data withholding and the next generation of scientists: Results of a national survey. Acad Med. 2006;81:128–136. https://doi.org/10.1097/00001888-200602000-00007

11. Milanovic F, Pontille D, Cambon-Thomsen A. Biobanking and data sharing: A plurality of exchange regimes. Genomics Soc Policy. 2007;3:17–30. https://doi.org/10.1186/1746-5354-3-1-17

12. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data sharing by scientists: Practices and perceptions. PLoS ONE. 2011;6, e21101, 21 pages. https://doi.org/10.1371/journal.pone.0021101

13. Shaw PM, Patterson SD. The value of banked samples for oncology drug discovery and development. J Natl Cancer Inst Monogr. 2011;42:46–49. https://doi.org/10.1093/jncimonographs/lgr004

14. Olson JE, Bielinski SJ, Ryu E, Winkler EM, Takahashi PY, Pathak J, et al. Biobanks and personalized medicine. Clin Genet. 2014;86:50–55. https://doi.org/10.1111/cge.12370

15. De Souza YG, Greenspan JS. Biobanking past, present and future: Responsibilities and benefits. AIDS. 2013;27:303–312. https://doi.org/10.1097/QAD.0b013e32835c1244

Big science and human development – what is the connection?

AUTHORS:
Michael Gastrow[1]
Thelma Oppelt[1]

AFFILIATION:
[1]Human Sciences Research Council – Education and Skills Development, Cape Town, South Africa

CORRESPONDENCE TO:
Michael Gastrow

EMAIL:
mgastrow@hsrc.ac.za

DATES:
Received: 03 Nov. 2017
Revised: 30 May 2018
Accepted: 12 July 2018
Published: 27 Nov. 2018

KEYWORDS:
economic development; innovation; science engagement; SKA telescope; large-scale science facilities

HOW TO CITE:
Gastrow M, Oppelt T. Big science and human development – what is the connection? S Afr J Sci. 2018;114(11/12), Art. #5182, 7 pages. https://doi.org/10.17159/sajs.2018/5182

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
DST-NRF Centre of Excellence in Human Development (South Africa)

The rationale for public expenditure and political support for large-scale science infrastructure is commonly underpinned by a universalist logic of big science's benefits. Literature assessing the impact of big science focuses on its contributions towards new fundamental insights about the universe; the development of skills, capabilities, networks, and innovation; and the development of globally transformative technology platforms that in turn make significant impacts on global human development. However, research into the local development impact of big science infrastructure is scarce. In this paper we reflect on the development impact of a big science project at the local level, drawing on the case study of the Square Kilometre Array telescope in South Africa's Karoo region. We find that the universalist logic that appears to apply at the global and national levels does not necessarily apply at the local level, where big science has resulted in human development benefits, but also substantial economic and social costs. On this basis we recommend that big science infrastructures, particularly in marginalised areas of developing countries, require a localised development proposition that takes into account local social complexities on the basis of extensive local engagement.

Significance:

- A synthetic review is presented of the different causal pathways through which big science may impact on human development.
- Analytical distinctions are developed between the human development impacts of big science at the global, national, and local scales.
- Considerations are put forward for a developmental agenda for big science facilities, particularly in developing countries.

Introduction

The presence of the Square Kilometre Array telescope (SKA) on the African continent has prompted reflection about the juxtaposition between large-scale globalised science and its host environment in South Africa's arid, sparsely populated and geographically isolated Karoo region. In a country where poverty, inequality and unemployment are serious social challenges, the human development implications of all policy choices, including science policy choices, are critical. To contribute to the national conversation about science policy and human development, including that related to the SKA, we present key findings from a literature review examining the relation between big science and human development. The literature review forms part of a project undertaken for the DST-NRF Centre of Excellence in Human Development. We apply general principles of this relation to the SKA, with the aims of gaining new insights into the manner in which this large-scale research infrastructure contributes to human development and reflecting on the big science–human development relation from a South African point of view.

The notion of 'big science' has emerged as shorthand for the increasingly large science projects that first proliferated during the Cold War.[1-3] The term was coined in 1961[4] to refer to a post-war political economy in which scientific research was a national security priority requiring state intervention and resources, as manifested for example in the first particle accelerators, as well as the Manhattan Project and much of the work of DARPA. The term 'large-scale science facilities' is also commonly used to refer to big science[5], primarily in relation to a facility's scale and role in systems of innovation. Nonetheless, in debates about human development impact, the two terms are broadly interchangeable.[5] 'Large-scale research infrastructure' has been seen to include telescopes; accelerators; facilities for biomedical research; sources for laser, neutron or synchrotron radiation, molecular imaging techniques, high magnetic fields, etc.; and high-performance supercomputers and grids.[5]

The rise of big science meant that scientists had to work in increasingly large hierarchical teams, and manage demands from new stakeholders: government, subcontractors and supply chains. Other analytical dimensions that distinguish big science from the broader scientific landscape include issues of geography (the growth of big science to encompass, sometimes literally, cities or regions), economics (the proliferation of multibillion dollar projects), multidisciplinarity (big science requires multiple academic disciplines and engineering technologies), and globalisation (big science projects generally require international collaboration at many levels).[2]

The knowledge and technology developed within big science facilities not only advance our understanding of the universe, but generate new classes of products and services that disrupt markets and change lives. At the same time, big science requires big funding, and hence policy trade-offs and public approval.[6,7] The scale of big science projects means that they exist in the public sphere – the notional space in which open debate takes place about matters of public interest.[8] In this public sphere, big science projects require social legitimacy, underpinned by public acceptance and mutual benefit. Such a social contract should ideally be supported by human development outcomes that justify the financial and opportunity costs of big science projects. Cultural, social and political contexts play important roles in shaping public perceptions of science.[9] For example, understanding public acceptance has been a key factor in informing the strategic decisions of nuclear fusion facilities, as it is for

the human genome project.[10-12] In both cases, public perceptions of the risks and benefits associated with big science play a central role in framing science projects as publicly acceptable or not. So, what is the relationship between big science and human development?

We began our analysis by examining some of the leading contemporary conceptions of human development, with a focus on Amaryta Sen's notion of 'development as freedom'[13], and Manual Castells' analysis of human development in the 'information age'[14], which position human development in relation to technological change, skills and knowledge, through a systemic actor-network analysis. On the basis of this conception of development, and its links to science and innovation, we examined the evidence and analysis of the relationship between big science and human development, and we put forward an abstracted model of the main causal pathways between them.

To explore the relationship between big science and human development in South Africa, our empirical focus shifts to South Africa's flagship science project, the SKA. We examined the (potential) human development implications of the SKA at the global, national and local levels. We explored the manner in which the human development implications of the SKA align with generalised conceptions of the big science–human development relationship, and the manner in which contextual factors determine a unique relationship at the local level. Through doing so, we aimed to extend the analysis of the role of big science in human development to include greater reflection about local impact, particularly for marginalised communities and developing countries.

Science, technology and human development

The theoretical debate over the nature of development is broad, and contains many strands which address different purposes and concepts.[15] In contrast to the study of economic development, which is focused on the dynamics of economic activity, the study of human development has a broader focus on the question of human well-being. This scope approximately aligns with the United Nations Sustainable Development Goals, and their aim of increasing prosperity and harmony in a sustainable world.

A suitable starting point for conceptualising human development is the consensus emerging from the 'intellectual coup' initiated by Sen[16], which re-framed the notion of human development as a question of access and capabilities, rather than economic resources. Sen[16] thus placed people at the centre of the development discourse. This conception of human development focuses on the cultivation of agency and capability, access to opportunity, and the freedom to work towards self-realisation in accordance with one's own beliefs and values. Development thus requires increasing levels of freedom, whether negative freedoms (freedoms from hunger, disease, poverty) or positive freedoms (freedom of self-expression, freedom of self-realisation). Sen[17] argues that freedom and capability are inextricably linked: freedom without the capability of its own realisation cannot contribute towards development. For example, famines may occur, not from lack of food, but because of a lack of resources to buy food. The freedom to achieve food security is meaningless without the capability to achieve it in practice. Sen's focus on capabilities and the development of human agency suggests that human development as an outcome of science and technology interventions should ideally be achieved through processes that include capability-building, social engagement and public participation.

For Castells, human development, also conceived as the freedom to exercise human capabilities, is tied to the production of resources and modes of social organisation that are increasingly dominated by technological drivers.[18,19] Castells conceives of these broad socio-economic dynamics as a shift into the 'global information age'. This age is defined as 'a historical period characterised by the technological revolution in information and communication, the rise of networking forms of social organisation, and the global interdependence of economies and societies'[20]. Castells views the organisation of infrastructures and activities for the creation, processing and transmission of information to be the main driver of wealth creation, a process termed 'informational development'.[19,20] The notion of informational development suggests that assessments of the relation between science, technology and human development include a focus on access to the Internet, to information and communication technology (ICT) more broadly, and to opportunities for technological upgrading, and their impacts on human development.

Castells' conception of the relationship between science and technology on the one hand, and human development on the other, is aligned with an innovation systems theory of wealth generation, in which value is created by 'transforming information into knowledge, and then applying knowledge to all the tasks to be executed on the basis of the technological and human capability embedded in the system'[20]. Innovation is commonly defined as the development of new products, processes or organisational structures.[21] Some strands of the innovation literature consider technology diffusion into households and the informal sector.[22] Innovation has a broad scope: a new product or process can be entirely novel, 'new to the world', or it can be 'new to the country', 'new to the sector', or 'new to the firm'.[21,23,24] The most common scale of analysis is at the national level, hence the term 'national systems of innovation' (NSI), defined as 'the network of institutions in the public and private sectors whose activities and interactions initiate, import, modify and diffuse new technologies'[22]. The NSI approach[25,26] remains widely used in the innovation studies literature[27], and has become the standard theoretical framework for guiding national and international science and technology policy[27,28].

Skills are fundamental enablers of innovation activity.[28] Without the requisite skills, new basic science cannot be performed, and new technologies cannot be developed, adapted or disseminated. The innovation studies literature suggests a strong causal interaction between the supply of higher levels of education, training and skills and increased demand for and supply of technical and organisational innovation.[29,30] Innovation and skills development are thus intertwined in an unfolding process that has been described as 'co-evolution'.[31,32] Co-evolution mechanisms include firm-level learning, technology and skills transfers through multinational corporations, local spillovers from innovation activities within firms, firm-level responses to the availability of local skills in terms of innovation activity, and university-firm interactions.[33,34]

Innovation systems analysis is largely concerned with understanding informational development as an input towards human development, rather than about human development per se. However, some strands of research within the discipline take the assessment of this relationship a step further. In his analysis of the linkages between techno-economic development and human development, Castells observed that despite the advancement of technologies, in particular ICTs, the majority of the global population remain in economically fragile and technologically excluded positions.[35] This concern also drove increased interest in 'innovation in inclusive development'.[21,27] This intersection between innovation systems analysis and development studies raises questions about the manner in which big science projects have, or have not, been inclusive and helped to drive social development. Key areas of investigation are the inclusiveness of the innovation process, the nature of participation among marginalised communities, and innovation in low-income and informal settings.[36]

Big science and human development

Given the significance of big science as a driver of technological change and knowledge generation, and the importance of human development as normative and policy goal, previous reviews of the intersection of these two domains have found a 'surprising lack of evidence on the nature and extent of the impacts of large-scale research facilities on the economy and society and on the mechanisms that generate such effects'[37]. Moreover, there is a lack of consensus in the literature over the reliability and generalisability of evidence about big science and innovation. The extant evidence does not encompass the full range of big science, as 'the evidence is skewed and cannot be extrapolated to the entire architectural and disciplinary diversity' of big science.[37] Although many researchers have made extensive claims about the effects of big science on innovation and economic growth, one review found 'insufficient evidence to support the claim that [big science

facilities] attract and retain talent and promote innovation', although 'more evidence exists that large infrastructures forge new networks and communities'[37]. However, the authors identified a lack of evidence, rather than a lack of impact – pointing towards the need for more research in this area in order to build an empirical base and move towards conceptual consensus.

These limitations notwithstanding, the logic underpinning analyses of the human development impact of big science is essentially 'universalist', supported by claims that the advances made by big science underpin most contemporary technological advancements, and through this advancement make an almost incalculable contribution to human development globally.[37] The rationale for public support for big science appeals to this universalist logic – according to which even the large financial costs incurred by basic science are outweighed by the (often unforeseen) benefits to humanity. At the same time, however, critiques of big science facilities have long questioned their return on investment, their proximity to military and industrial powers, and their opportunity cost in relation to scientific enterprises that have more direct benefits for human well-being.[4] The overall assessment of the contribution of big science to human development extends far beyond the scope of this paper. It may indeed be impossible to develop any meaningful quantitative assessment of the human development contribution of big science. However, the literature does establish a number of 'development logics' – notional processes through which big science facilitates or constrains human development – and these form the basis for our conceptualisation of a generic model of this relationship.

Figure 1 illustrates this model, showing the main causal pathways between big science and human development. We distinguish between the global and local scales, which are characterised by their inverse degrees of abstraction and specificity. The human development logic of big science is at its most abstract in the analysis of its global impact, whereas analyses of local human development impacts are largely determined by the specific social and economic context.

Firstly, as a backdrop to all human development considerations, the benefits of big science facilities need to be weighed against their financial costs, and the associated opportunity costs of dedicating resources to big science rather than more direct human-development-oriented interventions: 'Opportunity costs are at the heart of [the] issue'[37]. Trade-offs are made more difficult by the fact that the social benefits of scientific research are uncertain and take long to materialise.[38]

The logic of financial and policy trade-offs requires a clear conception of the benefits that will accrue with science expenditure. Such benefits are diverse, and of manifestly different types. Firstly, new knowledge contributes towards human development. New fundamental knowledge about reality has a normative value (understanding reality is a good in and of itself) and a positive value, as a more complete comprehension of the workings of the universe is *a priori* likely to have practical benefits too. Weighing up the nature of such benefits is difficult, as the normative value of new knowledge may vary across cultures, and the developmental impact of new fundamental knowledge is unpredictable and widely distributed across many economic and social formations.

Evidence and analysis are therefore largely focused on more material benefits. Some science facilities have an explicit social mission – for example biobanks or national reference laboratories, which are established to benefit public health, supported by scientific research. For such facilities, the social benefits are demonstrable, as they are tied to mission outcomes. However, most big science facilities exist for scientific purposes only, for example radio telescopes or particle accelerators: 'the challenge is to identify the societal footprint of facilities with a primary scientific mission'[37].

For such science facilities, innovation appears to be the primary mechanism through which human development outcomes are realised. Institutional assessments of big science projects, for example of CERN[39] and the European Extremely Large Telescope[40], and reviews by multilateral organisations such as the OECD[41], use an innovation systems framework, reflecting a primary interest in innovation, learning and economic impact, and focusing on technological and economic development as an output of interactions between actors in the system, and between systemic functions. Such institutional assessments have identified the main impacts of big science to be its influence on adjacent fields of science and technology, innovation incentives for industrial suppliers through procurement activity, and impact through community engagement and technology transfer.[42] A review undertaken by the OECD found that big science projects boost innovation and push knowledge-intensive work to world-class quality.[41] More specific applications of innovation theory include studies of how big science centres operate as learning environments for industrial supplier firms[42], as well as the broader mutual benefit of joint innovation processes between industry and big science[43]. A relatively extensive literature argues that social capital is the most significant factor in spillovers from big science into the private sector.[40,43]

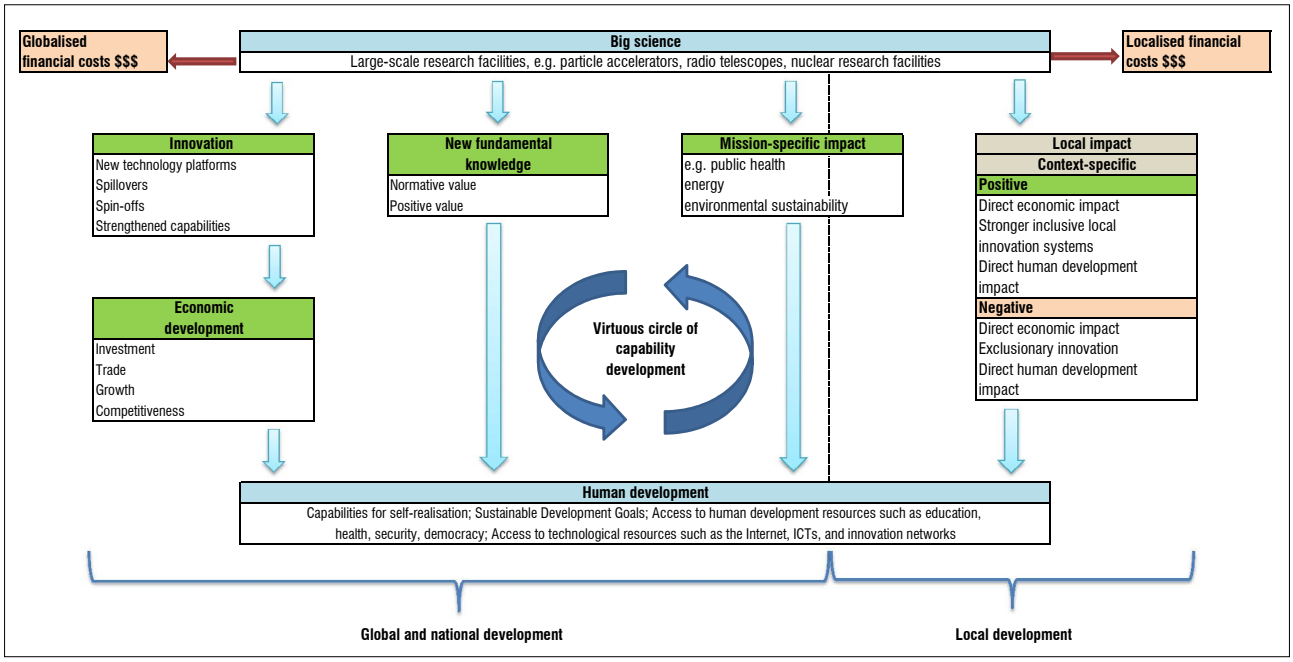


Figure 1: Big science and human development relationship in abstract.

Overall, taking into account the empirical limitations, our review identified a generalised pattern to the innovation benefits of big science: large research infrastructures attract the world's best researchers, on the basis of which the facility becomes a cluster or hub of knowledge in which there is heightened interaction between the scientific and the technological functions of the facility.[37] The interactions between big science centres, universities, industry and research institutes can be described as a virtuous circle, in which capabilities are strengthened through interaction, mutual learning and network building. As a consequence, there is consensus in the literature that big science provides a positive return on investment, and is of substantial net benefit to society[37,44], even taking into account the large financial and opportunity costs.

The technologies developed for big science facilities are diffused into other sectors of society, creating economic benefits for innovators and social benefits for society.[45] Related to these benefits is the potential of commercial spin-offs based on new technologies and capabilities developed for science projects.[46] Research-based spin-offs are small, new technology-based firms whose intellectual capital originated in universities or other public research organisations. Innovation spillovers occur when an organisation stimulates external technological improvements through internal innovation. Such external benefits often outweigh the initial investment in big science, although such benefits are not immediate.[47] The benefits of spin-offs are partly conjectural at the beginning of a big science project, but may emerge as their main contribution to society by maturity of the project.

Big science facilities, along their path of pursuing new fundamental knowledge, have pushed the boundaries of technology, and through this enabled the establishment of technological platforms such as the Internet, Wi-Fi, the transistor, nuclear energy, GPS, touch screens, genetic medicine, and so on. Such advances have historically been generally serendipitous, for example, the development of the touch screen and the World Wide Web at CERN[41], and the origins of the Internet within the USA's National Physics Laboratory and associated supercomputer systems[48]. The net contribution of such technological platforms to human progress is indeed incalculable, and can perhaps best be understood as diffuse global drivers of technological change, and all the attendant impacts on development.

Studies that focus on the social and cultural effects of big science are far less numerous than those using an innovation lens. Previous reviews have pointed out a general paucity of studies looking at the social impact of big science.[5] Extant research is often indirect and more diverse in its theoretical perspectives; for example, the sociology of knowledge creation in big-science contexts[49], using social constructivism theory to examine impacts of large-scale research facilities[50], and using grounded theory together with theoretical constructs from social network, social capital, and inter-organisational learning theories[42]. A broad overview of the social impacts of big science, in which seven case studies of large-scale science facilities were explored, showed that 'the least understood, yet potentially the most significant aspect of big science facilities operates through their broader contributions to society and culture'[5].

Whereas global and national scales allow for an abstraction of the relationship between big science and human development, at the local scale the relationship is more context specific, may include both significant positive and negative impacts, and may fall outside the techno-economic scope that forms the focus of assessments at national and global levels. The few studies that focus on local impact are located in developed countries, and also focus on the innovation and economic mediating processes, rather than human development as an ends. For example, a study of the impact of the Fermilab on the city of Chicago and the State of Illinois[51] found that direct expenditure boosted the local economy by USD288 million, and created 4500 jobs which led to net earnings to households and businesses in the region of USD643 million. The human development impact, for example as related to the Sustainable Development Goals, was not assessed. However, further research into the human development impact of the new employment and income would conceivably reveal significant benefits.

Not all big science projects experience harmonious relationships with their host publics. To use an example in the astronomy domain, the Thirty Metre Telescope in Hawaii, while under construction and on course to become a flagship international instrument, had its construction permit revoked on the grounds that public consultation processes had not been followed.[52,53] Mauna Kea is the most sacred mountain in Hawaiian culture, and one of the key objections was that the instrument would be built on a sacred ancestral burial ground. In this case, negative local impacts led to a breakdown in legitimacy that presented a material risk to the project. Importantly, the local developmental cost was unrelated to scientific, innovation or economic outcomes. Instead the objection related to local participation, agency and culture – issues that are aligned with Castells' notion of agency and participation being central to the relationship between science, technology and human development.

The Square Kilometre Array and human development

The SKA, currently under construction, will become the world's largest telescope. In South Africa and its eight African partner countries, the high- and mid-frequency radio receiver array will include approximately 2000 15-metre-wide receiver dishes, all linked via fibre-optic cable and supercomputers. In Australia, the low frequency array will consist of tens of thousands of small antennae dipoles. The SKA's primary science questions are related to the origins of the universe, the formation of the first stars and galaxies, the nature of dark matter and dark energy, theories of gravity and relativity, and the search for extra-terrestrial life.

The project has a global set of stakeholders, funders and employees, and a globalised innovation network.[54] The SKA has distinct implications for the two main host countries, South Africa and Australia, African partner host countries (Botswana, Ghana, Kenya, Madagascar, Mauritius, Mozambique, Namibia and Zambia), the headquarter country (the UK), and the eight other countries in the global SKA consortium (Canada, China, India, Italy, New Zealand, Sweden, the Netherlands and the UK). Conceptualisation of the SKA's relation to human development therefore needs to span the global and the local, the immediate and the long term. Within the South African context, the Karoo region – including the towns of Carnarvon, Williston, Loxton and Brandvlei, as well as surrounding farming communities – forms an important localised setting for assessing the impact of the SKA. The impact in urban centres such as Cape Town and Gauteng is also significant because of the location of the SKA's central and regional offices. The impact on universities, other astronomy facilities, and South Africa's broader high-skills economy, are all potential areas for investigation.

The SKA provides a promising case study for assessing the big science–human development relationship, and in particular may add to the literature addressing this question in Africa and in developing countries. To guide this assessment, we examined the SKA in terms of its adherence to the postulated generic model of big science and human development, and examined the local context to explore how this relationship has been manifested in South Africa's Karoo region. Within this framework, key analytical questions that emerged from our conceptualisation of human development and innovation included those related to access to ICT, opportunities for technological upgrading, the inclusiveness of the innovation process, the nature of participation among marginalised communities, and evidence of movement towards human development goals such as increased health, literacy and employment.

The core knowledge and financial parameters of the SKA at the global level align with the generalised model of big science and human development. The cost of the international project[55] is estimated at EUR1.8 billion, representing a significant opportunity cost, and a requirement for funders to provide substantiation to their stakeholders and constituencies. The expenditure will be distributed globally across the two infrastructure sites in Australia and South Africa, in South Africa's eight African partner countries, the headquarters in the United Kingdom, and throughout the SKA's global supply chain and innovation network.[54] The South African government is contributing significantly to the financing of the telescope:

ZAR2.3 billion has been set aside for the 2018 Medium Term Expenditure Framework, which covers a 3-year period.[56]

The generation of new fundamental knowledge by the SKA, like other big science projects, has the potential to revolutionise our conception of the universe and our place therein. Progress in even one of its six main science objectives would enormously advance our understanding of physics and cosmology. Technological breakthroughs could also have enormous global benefits. It is conceivable that advances in the areas of big data, supercomputing, and algorithm development may generate new technological platforms with a global impact and widespread implications for human development. However, it is difficult, if not impossible, to foresee what such implications might be.

At the national level, the impact of the SKA on technological capabilities and the NSI may prove to be its greatest, albeit indirect, contribution to human development in South Africa. A NSI with enhanced capabilities can compete more effectively in the global knowledge economy, and play a greater role in economic growth and development. Building knowledge transfer capabilities by enhancing coordination, interaction and alignment between skills supply and skills demand is an important contribution of the SKA to the NSI. These contributions include mechanisms such as the Universities Working Group, which informally coordinates skills development activities at South African universities; participation in the National Astrophysics and Space Sciences Programme Steering Committee, which coordinates postgraduate curricula at the universities; and SARChI Chairs in various science and engineering focus areas relevant to the SKA, which bring in international expertise as well as local expertise, and draw on this expertise to develop new knowledge and skills among postgraduate students and postdoctoral positions.[54]

Technology diffusion and technological spin-off activity is another manner in which the SKA strengthens the NSI and contributes towards economic growth. Technological capabilities built within the SKA find their way into other sectors, particularly those that utilise big data, advanced ICTs, and engineering – although the nature of such spillovers has not been subject to research. In terms of spin-off activity, the SKA has a dedicated office to support this function, which is seen as important to its public benefit role in the NSI and to strengthening its own financial sustainability. Private-sector spin-off firms cultivate new economic activity by participating in high-tech global value chains.[54] However, the nature and extent of these effects have not been comprehensively researched, pointing the way towards a future research agenda in this area.

Given the significance of the co-evolution of skills and innovation capabilities, a central aspect of the SKA's contribution in South Africa has been its Human Capital Development Programme (HCDP), which aims to develop the capabilities of engineers, scientists, technicians and artisans to build, operate and use the SKA. At the inception of the HCDP in 2005, South Africa was home to only a handful of radio astronomers, and had limited technical capabilities – while competing against an internationally competitive and much larger Australian radio astronomy sector. Today, the sector has grown by orders of magnitude: as of 2017, the HCDP had awarded 943 grants to support postdoctoral fellows, postgraduate and undergraduate students, and students training to be artisans, and had also funded five dedicated university-based Research Chairs. The HCDP is funded by the Department of Science and Technology, and from 2005 to 2017 had cost a total of ZAR446 million. (Data provided by the SKA.)

Overall, therefore, the evidence suggests that, at the national level, the generalised model of the human development benefits of big science largely applies to the SKA too. However, at the local level, contextual factors become more significant, and the balance between benefits and costs is shifted. The areas surrounding the SKA's core site in the Karoo are where the project has its greatest direct exposure to local communities. The SKA has rendered substantial economic benefits to local communities. As of 2016, the organisation's activities in the Northern Cape Province have included ZAR136 million spent through local suppliers, and the creation of 7284 employment opportunities.[57] In addition, indirect economic benefits include a growing market for accommodation and tourism.[58] The local hospitality industry has benefitted, as visitors to the SKA must largely stay off-site, and find accommodation in surrounding towns – although these effects have yet to be measured.

In addition to the SKA's economic impact, the HCDP intersects with local communities through its schools programme. As of 2016, SKA had provided training to 351 people from Northern Cape communities, rolled out a support programme for eight local schools, involving more than 4000 learners, and funded 9 local students to attend university and 72 students to attend vocational colleges. Carnarvon High School has benefitted from a new computer lab, high-bandwidth Internet connection, bursaries, and direct engagements with the SKA. The SKA has also routed some resources towards direct human development interventions, on the basis of engagement with local community stakeholders. For example, the SKA has sponsored the Carnarvon Library, and is providing small-scale funding to a non-profit organisation that undertakes skills training, foetal alcohol syndrome awareness and intervention programmes, arts and crafts at the local high school, and a reading programme at the primary school.

However, there are also a range of economic costs to local communities that have resulted from the SKA. The purchase of farming land to host the telescope has removed the associated economic activity from local supply chains, negatively impacting on both upstream and downstream sectors. Examples include reduced demand for agricultural supplies bought from local firms, and reduced demand for services of the Carnarvon abattoir. The Strategic Environmental Assessment of the Council for Scientific and Industrial Research[59] developed an economic model that predicted a loss in annual agricultural production of ZAR16 million, a loss in throughput at abattoirs equivalent to 8.24% of annual slaughter volumes, a drop in sales and business volumes amongst local business and entrepreneurs equivalent to ZAR9.09 million, and a loss in production value on neighbouring farms adjacent to the SKA core because of increased predator activity. Cumulatively, these effects would lead to an estimated economic loss of ZAR31 million annually, and a cumulative loss of 1565 jobs.

While the SKA has clearly contributed to South Africa's NSI, there is currently little evidence to suggest that local innovation systems in the Karoo have been strengthened. This question is a subject of current research being conducted by the Human Sciences Research Council for the National Research Foundation. The SKA core site, for technical reasons, remains closed to the public, and the advanced technologies deployed at the site have little or no impact in surrounding communities. Attempts to directly link to local systems of innovation may in any case be ineffective, given the large gap between the SKA's advanced technologies and local capacities to absorb new technologies. Moreover, the radio telescope has significantly constrained Internet access around its core site in the Karoo, negatively impacting on ICT access for some farmers and surrounding communities.

Castells' agent-centric notion of human development argues that social engagement and public participation should be central to science that seeks to contribute towards human development. This argument raises the question of the extent and efficacy of the SKA's engagement with local communities. The evidence here is mixed, revealing a wide range of engagement activities, as well as ongoing local dissatisfaction with engagement mechanisms. Local civic resistance to the SKA has catalysed around claims of insufficiently mitigated economic losses to special interest groups, and claims of inadequate local agency in the telescope's activities in the Karoo.[60] An open forum for community engagement was in 2016 shut down as a result of local opposition to the SKA[61], and the SKA's primary local engagement platform has since then been through local municipalities. The socio-political dynamics of the interactions between civil society formations and the SKA is an area of ongoing research, including assessments of social media as a platform for resistance.[62]

At the local level, therefore, the relationship between big science (manifested in the SKA) and human development does not adhere to the same developmental logic that applies to the global level. There are immediate economic benefits and economic costs, and the causal path from big science to human development is more direct, rather than

being mediated through complex innovation and economic systems. The new fundamental knowledge that may emerge from the telescope is of little current value to the local population, and local innovation systems have been minimally affected by the telescope. Instead, the main human development outcomes have been related to dedicated development initiatives, such as education interventions and social development interventions. The effects on the most pressing human development needs – poverty, unemployment, alcoholism, foetal alcohol syndrome, and other social problems – appears to be small, although it has yet to be measured; no impact assessments of the SKA's education and social interventions have yet been published.

Reflections on big science and the developmental context

The scale of big science opens up questions of scale for social scientists too: we need to define and disentangle the effects that big science has on the world, on nations, and on regions surrounding science infrastructure. It would be an error to see the effects at these different scales as being directly comparable – they differ enormously in terms of their economic scale and scope, time frames, predictability, measurability, causal mechanisms and policy context. The potential effects on humanity of new fundamental insights into the nature of reality, or the emergence of a new disruptive technological platform, cannot be meaningfully compared, in any direct sense, to the immediate impact of social development interventions surrounding a scientific facility. Nonetheless, the rationale for undertaking big science, and funding big science, must in some way take all these diverse effects into account, and proceed on the basis of a clear conceptualisation of human development impact that provides a balanced development proposition at all levels.

At the global and national levels, a broadly conceived consensus appears to exist (empirical limitations notwithstanding) with respect to the development logic underpinning claims about big science's contribution to global and national human development, firstly by placing value on the development of new fundamental knowledge in and of itself, and secondly through innovation as a mediating process that translates big science activities into economic and social benefits. However, this logic is less clearly defined at the local scale, where the economic and social benefits and costs are not mediated through complex global innovation and production systems, but rather result as a direct effect of a facility's social mission, or through employment, infrastructure, supply chains, local innovation systems and direct development interventions. At the local level, participation and agency become critical factors in terms of achieving a developmental process that aligns with conceptions of agency-based human development, as well as establishing social legitimacy and a 'licence to operate'.

In the case of the SKA, arguments for the extensive global and national benefits of the project are not relevant to local stakeholders, who are affected by immediate and local consequences. Extant evidence suggests that the local economic impact has, on balance, been positive, although special interest groups, such as the agricultural sector, may experience negative consequences. Capabilities have certainly been built, particularly through the SKA's schools programme. However, many of the innovation benefits of the SKA, including knowledge spillovers, economic spin-offs and capability development, occur in urban centres but not in the Karoo, showing that the postulated strengthening of innovation systems through big science may exclude marginalised groups and by-pass local systems of innovation. Local resistance in some instances frames local communities as lacking agency with respect to decisions about SKA, and as paying the true cost for the instrument.

These disjunctures suggest several imperatives for big science infrastructures, particularly those being built in developing countries and marginalised regions. Firstly, there is a need to balance global, national and local human development processes, priorities and outcomes. There is a need to put forward a development proposition that provides clear human development benefits at all levels. There is a need to include local stakeholders in decision-making processes. And there is a need to present a clear argument that global and national benefits, as enormous as they may be, are not being prioritised over local costs.

Acknowledgements

The support of the DST-NRF Centre of Excellence in Human Development is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the Centre of Excellence in Human Development. We thank Prof. Linda Richter and Dr Vijay Reddy for their valuable contributions at various stages of the research. We also thank all the SKA staff and members of the Carnarvon and Loxton communities who participated in the research.

Authors' contributions

M.G.: conceptualisation, project leadership, project management, funding acquisition, methodology, analysis, writing initial draft, writing revisions. T.O.: conceptualisation, data collection, literature review.

References

1. Kaiser D. From blackboards to bombs: Seventy years after the destruction of Hiroshima and Nagasaki by nuclear weapons, David Kaiser investigates the legacy of 'the physicists' war'. Nature. 2015;523(7562):523–526. https://doi.org/10.1038/523523a
2. Galison P, Hevly B, Weinberg AM. Big science: The growth of large-scale research. Phys Today. 1992;45:89. https://doi.org/10.1063/1.2809880
3. Sullivan WT, editor. The early years of radio astronomy: Reflections fifty years after Jansky's discovery. Cambridge: Cambridge University Press; 2005.
4. Weinberg AM. Impact of large-scale science on the United States. Science. 1961;134(3473):161–164. https://doi.org/10.1126/science.134.3473.161
5. Autio E. Innovation from big science: Enhancing big science impact agenda [document on the Internet]. c2014 [cited 2018 May 30]. Available from: http://data.parliament.uk/DepositedPapers/Files/DEP2014-0843/Innovation-from-big-science-enhancing-big-science-impact-agenda.pdf
6. National Audit Office. Big science: Public investment in large scientific facilities. Norwich: The Stationery Office; 2007. Available from: https://www.nao.org.uk/report/big-science-public-investment-in-large-scientific-facilities
7. Cost.eu. Benefits of research infrastructures beyond science – The example of the Square Kilometre Array (SKA); 2010 March 30–31; Rome, Italy. Available from: http://www.cost.eu/events/ska
8. Habermas J. The structural transformation of the public sphere: An inquiry into a category of bourgeois society. Cambridge, MA: MIT Press; 1991.
9. Reddy V, Gastrow M, Juan A, Roberts B. Public attitudes to science in South Africa. S Afr J Sci. 2013;109(1–2), Art. #1200, 8 pages. http://dx.doi.org/10.1590/sajs.2013/1200
10. Bauer MW. Public perceptions and mass media in the biotechnology controversy. Int J Public Opin Res. 2005;17(1):5–22. https://doi.org/10.1093/ijpor/edh054
11. Prades López A, Horlick-Jones T, Oltra C, Solá R. Lay perceptions of nuclear fusion: Multiple modes of understanding. Sci Public Pol. 2008;35(2):95–105. https://doi.org/10.3152/030234208X282853
12. Cunningham-Burley S, Kerr A. Defining the 'social': Towards an understanding of scientific and medical discourses on the social aspects of the new human genetics. Sociol Health Illn. 1999;21(5):647–668. https://doi.org/10.1111/1467-9566.00177
13. Sen A. Development as freedom. Oxford: Oxford University Press; 1999. http://dx.doi.org/10.2307/40203469
14. Castells M, Himanen P, editors. Reconceptualizing development in the global information age. Oxford: Oxford University Press; 2014. https://doi.org/10.1093/acprof:oso/9780198716082.001.0001
15. Pieterse JN. Development theory. London: Sage; 2009. http://dx.doi.org/10.4135/9781446279083
16. Sen AK. Equality of what? In: McMurrin S, editor. Tanner lectures on human values. Cambridge: Cambridge University Press; 1995. p. 197–220. https://doi.org/10.1093/0198289286.003.0002
17. Sen AK. Development as capability expansion. In: DeFilippis J, Saeger S, editors. The community development reader. 2nd ed. New York: Routledge; 2012. p. 319–327. https://doi.org/10.1007/978-1-349-21136-4_3

18. Castells M. Materials for an exploratory theory of the network society. Br J Sociol. 2000;51(1):5–24. https://doi.org/10.1080/000713100358408
19. Castells M, Himanen P. The information society and the welfare state: The Finnish model. Oxford: Oxford University Press; 2002.
20. Castells MH. Reconceptualizing development in the Global Information Age: Oxford: Oxford University Press; 2014.
21. Martin BR. The evolution of science policy and innovation studies. Res Policy. 2012;41(7):1219–1239. https://doi.org/10.1016/j.respol.2012.03.012
22. Freeman C. The 'National System of Innovation' in historical perspective. Cambridge J Econ. 1995;19(1):5–24.
23. Nelson R, Winter S. An evolutionary theory of economic change. Cambridge, MA: Belknap Press; 1982.
24. Mowery D, Rosenberg N. Technology and the pursuit of economic growth. Cambridge: Cambridge University Press; 1991.
25. Lundvall BÅ. National systems of innovation: Towards a theory of innovation and interactive learning. London: Pinter; 1992.
26. Nelson RR, editor. National innovation systems: A comparative analysis. Oxford: Oxford University Press; 1993.
27. Lundvall BÅ. Innovation system research and policy. Where it came from and where it might go. Paper presented at: InCAS Seminar; 2007 December 04; Oslo, Norway.
28. Fagerberg J, Verspagen B. Innovation studies – The emerging structure of a new scientific field. Res Policy. 2009;38(2):218–233. https://doi.org/10.1016/j.respol.2008.12.006
29. Toner P. Workforce skills and innovation: An overview of major themes in the literature. OECD Education Working Papers. 2011;(55):0_1.
30. Cohen WM, Levinthal DA. Absorptive capacity: A new perspective on learning and innovation. Adm Sci Q. 1990;35(1)128–152. https://doi.org/10.2307/2393553
31. Lewin AY, Couto V. Next generation offshoring: The globalization of innovation: 2006 survey report. Durham, NC: Centers for International Business Education and Research (CIBER), Duke University; 2007.
32. Losby B. Knowledge, institutions and evolution in economics. London: Routledge; 2000.
33. Metcalfe JS. Evolutionary economics and creative destruction. London: Routledge; 1998. https://doi.org/10.4324/9780203275146
34. Lorentzen J. MNCs in the periphery: DaimlerChrysler South Africa (DCSA), human capital upgrading and regional economic development. In: Multinationals on the periphery. London: Palgrave Macmillan; 2007. p. 158–187. https://doi.org/10.1057/9780230593046_7
35. Castells M. End of millennium. Vol. III: The information age: Economy, society and culture. Oxford: Oxford University Press; 1998.
36. Cozzens S, Sutz J. Innovation in informal settings: A research agenda. Ottawa, Canada: International Development Research Centre; 2012.
37. Horlings E, Gurney T, Somers A, Van den Besselaar P, Van Saksenlaan A. The societal footprint of big science. Rathenau Instituut working paper. Den Haag: Rathenau Instituut; 2012. Available from: http://hdl.handle.net/20.500.11755/16c47732-84e4-4a9a-8c2f-be53d12eade4
38. Martin BR, Tang P. The benefits from publicly funded research. Brighton: Science Policy Research Unit, University of Sussex; 2007.
39. Autio E, Hameri AP, Bianchi-Streit M. Technology transfer and technological learning through CERN's procurement activity. Geneva: CERN; 2003.
40. Science and Technology Facilities Council (STFC). E-ELT impact: The impact of the European Extremely Large Telescope. Edinburgh: STFC; 2009. Available from: www.eelt.org.uk/astronomers/elt-impact.pdf
41. Organisation for Economic Development and Co-Operation (OECD). The impacts of large research infrastructure on economic innovation and on society: Case studies. Paris: OECD; 2014. Available from: http://www.oecd.org/sti/sci-tech/CERN-case-studies.pdf
42. Autio E. Innovation from big science: Enhancing big science impact agenda. London: Department of Business, Innovation & Skills, Imperial College Business School; 2014.
43. Autio E, Hameri AP, Vuola O. A framework of industrial knowledge spillovers in big-science centres. Res Policy. 2004;33(1):107–126. https://doi.org/10.1016/S0048-7333(03)00105-7
44. Vuola O, Hameri AP. Mutually benefiting joint innovation process between industry and big-science. Technovation. 2006;26(1):3–12. https://doi.org/10.1016/j.technovation.2005.03.003
45. Hickling Arthurs Law Corporation. Return on investment in large scale research infrastructure. Ottawa: National Research Council; 2013.
46. Gómez AL. Technological spillovers of research infrastructures. Departmental Working Papers 2015-18, Department of Economics, Management and Quantitative Methods at the University of Milano. 2015. Available from: http://wp.demm.unimi.it/files/wp/2015/DEMM-2015_18wp.pdf
47. Byckling E, Hameri AP, Pettersson T, Wenninger H. Spin-offs from CERN and the case of TuoviWDM. Technovation. 2000;20(2):71–80. https://doi.org/10.1016/S0166-4972(99)00113-3
48. Rizzuto C. Benefits of research infrastructures beyond science. Presented at: ERF Workshop on the Socio-Economic Relevance of Research Infrastructures; 2012 May 31 – June 01; Hamburg, Germany.
49. Yates DM. Turing's legacy: A history of computing at the National Physical Laboratory 1945-1995. London: Science Museum; 1997. p. 126–146.
50. Knorr CK. Epistemic cultures. How the sciences make knowledge. Cambridge, MA: Harvard University Press; 1999.
51. Sallee CM, Watkins SD, Rosaen AL. The economic impact of Fermi National Accelerator Laboratory. Anderson Economic Group report to the University of Chicago [document on the Internet]. c2011 [cited 2018 May 30]. Available from: https://ovprnl.uchicago.edu/sites/research.uchicago.edu/files/Fermilab_Economic_I mpact_Full_Study.pdf
52. Sanders GH. The thirty meter telescope (tmt): An international observatory. J Astrophys Astron. 2013;34(2):81–86. https://doi.org/10.1007/s12036-013-9169-5
53. Overbye D. Hawaii court rescinds permit to build thirty meter telescope. New York Times. 2015 December 03. Available from: https://www.nytimes.com/2015/12/04/science/space/hawaii-court-rescinds-permit-to-build-thirty-meter-telescope.html?_r=0
54. Gastrow M. Understanding interactive capabilities for skills development in sectoral systems of innovation: A case study of astronomy and the Square Kilometre Array telescope. LMIP report 6. South Africa: Labour Market Intelligence Partnership; 2015. Available from: http://www.lmip.org.za/sites/default/files/documentfiles/HSRC%20LMIP%20Report%206%20Web_0.pdf
55. Frequently asked questions about the SKA [webpage on the Internet]. No date [cited 2018 May 30]. Available from: https://www.skatelescope.org/frequently-asked-questions/
56. National Treasury. Estimates of national expenditure 2018. Pretoria: National Treasury; 2018. Available from: http://www.treasury.gov.za/documents/national%20budget/2018/enebooklets/Vote%2030%20Science%20and%20Technology.pdf
57. Adam R. SKA SA's investment impact on the Northern Cape. SKA Resources. 2017. Available from: http://www.ska.ac.za/wp-content/uploads/2017/03/ska_investment_northern_cape_2017.pdf
58. Ingle M. Making the most of 'nothing': Astro-tourism, the sublime, and the Karoo as a 'space destination'. Transform Crit Perspect South Afr. 2010;74(1):87–111. https://doi.org/10.1353/trn.2010.0013
59. Council for Scientific and Industrial Research (CSIR). Strategic environmental assessment for the South African mid-frequency array of SKA Phase 1. CSIR report CSIR/02100/EMS/ER/2016/15240/B. Pretoria: CSIR; 2016.
60. Butler SS. Knowledge relativity: Carnarvon residents' and SKA personnel's conceptions of the SKA's scientific and development endeavors. Stellenbosch: Stellenbosch University; 2018.
61. Wild S. Giant SKA telescope rattles South African community. Nature. 2016;534:444–446. https://doi.org/10.1038/534444a
62. Binneman A. The SKA's struggles with anti-SKA advocacy groups. Presented at: Public Communication of Science and Technology Conference; 2018 April 03–06; Dunedin, New Zealand.

Perverse incentives and the political economy of South African academic journal publishing

AUTHOR:
Keyan G. Tomaselli[1]

AFFILIATION:
[1]Department of Communication Studies, University of Johannesburg, Johannesburg, South Africa

CORRESPONDENCE TO:
Keyan Tomaselli

EMAIL:
Keyant@uj.ac.za

DATES:
Received: 29 Jan. 2018
Revised: 02 Feb. 2018
Accepted: 24 July 2018
Published: 27 Nov. 2018

KEYWORDS:
rent-seeking; publication subsidy; cultural economy; South Africa; research incentives

HOW TO CITE:
Tomaselli KG. Perverse incentives and the political economy of South African academic journal publishing. S Afr J Sci. 2018;114(11/12), Art. #4341, 6 pages. https://doi.org/10.17159/sajs.2018/4341

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
None

Academic publishing in South Africa attracts a state research incentive for the universities to which the authors are affiliated. The aim of this study was twofold: (1) to examine the composition of the research value chain and (2) to identify the effects of broken links within the chain. The methodology selected was a lived cultural economy study, which was constructed through incorporating dialogue with editors, authors and researchers in terms of my own experience as a journal editor, read through a political economy framework. The prime effect is to exclude journals, especially independent titles, from directly earning publishing incentives. The behaviour of universities in attracting this variable income is discussed in terms of rent-seeking which occurs when organisations and/or individuals leverage resources from state institutions. Firstly, this process commodifies research and its product, publication. Secondly, the value chain is incomplete as it is the journals that are funding publication rather than – in many cases – the research economy funding the journals. Thirdly, authors are seeking the rewards enabled by the incentive attached to measurement systems, rather than the incentive of impacting the discipline/s which they are addressing. Fourthly, the paper discuses some policy and institutional matters which impact the above and the relative costs between open access and subscription models. Editors, journals and publishers are the un- or underfunded conduits that enable the transfer of massive research subsidies to universities and authors, and, in the case of journals, editors' voluntary work is the concealed link in the value chain enabling the national research economy.

Significance:

- The South African scientific publishing economy is built on a foundation of clay: this economy distorts research impact and encourages universities and academics to commoditise output.

Introduction

Amongst the items usually under discussion in academia are peer review and the alleged unreasonable profits made by multinational publishing firms.[1] Allied to these are issues of open access.[2] Finally, the question of predatory journals that prey on the 'publish or perish' syndrome is a growing concern.[3-6] These four topics background more specific concerns addressed here regarding the issue of incentive-seeking by South African universities, enabled by the unique economics of academic journal publishing in South Africa. This uniqueness is illustrated in the final section which draws on the textured experience of a number of editors and production editors based in South Africa who have commented, some multiple times, on earlier drafts of this analysis. My conclusion is that the business of academic journal publishing has, as a consequence of the way in which research is funded from the public purse, become the business of subsidising aspects of the business of education.

South African universities earn substantial rewards from the Department of Higher Education and Training (DHET) when their affiliates (academics, students and honorary appointments) publish in the journals that comprise 'the DHET-accredited list'. The benefit of this DHET incentive mechanism has been a significantly increased publication output and amplified productivity, through encouraging more academics to engage in research and publication. However, the negative outcome is that this system drives many universities to become rent-seeking, or at minimum, engage in the pursuit of perverse incentives. Rent-seeking is:

> *The process whereby organisations or individuals expend resources to obtain actions from state institutions that allow these actors to earn 'rents' in excess of what they would earn in the hypothetical scenario of a competitive market.*[7]

Allied to my use of rent-seeking is the term 'perverse incentive', that is, an unintended and sometimes undesirable outcome that contravenes the intention of the incentive's designers, in this case the state's policymakers. Incentive payments for publication in journals is unique to South Africa.

The hard data identifying incentive-seeking behaviour with regard to (1) universities, (2) specific journals and (3) even specific authors, has been summated in omnibus quantitative surveys conducted by the Centre for Research on Evaluation, Science and Technology (CREST, Stellenbosch University) for the Academy of Science of South Africa (ASSAf).[6] My analysis draws directly on this hard data.

Incentive-seeking behaviour involves effort by private interests to capture excess rent/surplus by influencing the state's use of its power. Social harm is thereby caused in two main ways, the first of which is through distortion in the allocation of resources, the details of which depend on what precisely the rent-seeking concerns. The second is through the costs incurred by those private (in our case, public universities) interests in seeking to secure this outcome.[8,9] DHET's financial incentive scheme encourages publication through the way that resources are transferred to universities. This incentive occurs in the state's generation of a new form of 'surplus' – one

that typifies the knowledge economy in which information becomes an intangible commercial good – that can be pursued by researchers. The rules for accessing that surplus are weak, as will become evident below.[6,7] The use of extrinsic incentives, however, leads to rent-seeking behaviour at lower levels. The publication churn in predatory or low-quality journals is one response to the incentive. Another is the policy adopted by some institutions to allocate a portion of the incentive directly to individuals as taxable income. Anecdotal evidence further reveals corrupt agreements by some individuals who have deliberately published in known predatory, but nevertheless accredited titles, like the *Mediterranean Journal of Social Sciences*.[4,7]

The system's design exposes it to abuse and fails to address intellectual or academic quality in a satisfactory way, other than via ASSAf's 5-yearly journal assessments (see below). Kerr and de Jager[5] and Mouton[6] have shown that some journals on the DHET list (mainly listed on the ProQuest International Bibliography of Social Sciences) have been also classified as 'predatory' on Beall's List (http://beallslist.weebly.com).

Apart from the above-discussed unintended consequences flowing from the application of the DHET incentive system by individual universities, the scheme is structurally flawed because it resources only some links of the value chain. A value chain links the complete range of actions – in the case of journals, editorial, peer review, design, production, marketing and distribution – that combine to deliver a product or service. In academic research, the value chain starts with the raw material (research data) used by researchers to make their products (articles), and includes all add-ons prior to the published work being sold to markets (libraries, students, academics, researchers, policymakers, etc.) (Figure 1).

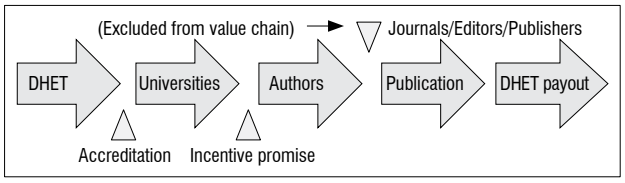


Figure 1: The journal publishing value chain.

The supply of articles and the purchase of journal subscriptions are resourced within the value chain. Unresourced within the chain, however, are editorial, peer review, production and publishing processes. Only two journal editors questioned the ideological motivation of the DHET system when it was first introduced in the early 1980s, but they both missed the financial basis of the initiative where non-state-funded journals were concerned – an omission that has continued into the current post-apartheid period. These were the editors of *Critical Arts* (Tomaselli et al.[10]) and *Scrutiny2* (Diedre Byrne[11]). The system is now driven by neoliberal productivity imperatives rather than apartheid ideology. That is to say, the editors of these two journals had argued that the apartheid Department of Education's incentive system was designed to shape research outcomes in its favour by accrediting specific journals. The application process at that point was opaque and the selection criteria of journals was unknown, as was the entity or committee or office that actually conferred 'accreditation'.

As the financial support of peer-review, editorial and production costs continue to be excluded by the value chain, the system imposes burdens on editors and peer reviewers in terms of article oversupply. This oversupply is not only a natural consequence of increasing research productivity, but also a consequence of rent-seeking. These outcomes are related: the research economy is characterised by post-publication resources (the incentives) generating production capacity on the one hand and the oversupply of submissions (as a kind of pre-product) on the other hand. This asymmetry is analysed here through examining (1) the political economy of the national research publication system, specifically (2) the exploitation of voluntary editors and reviewers; and identifying (3) instances of perverse incentive-seeking by universities which leverage journals and authors (including graduate students) as 'cash cows'.

Accredited journals

South African university administrations are obsessed with their faculty publishing in so-called 'accredited journals'. From an unknown selection process during apartheid, an accredited journal now meets clear publicised and monitored technical and editorial criteria determined by DHET. Journals indexed on Clarivate Analytics' (previously Thomson Reuters') Web of Science, the ProQuest International Bibliography of Social Sciences (IBSS) and Scopus are accorded honorary accreditation. The inclusion of the Norwegian list is not mentioned here as this choice has never been satisfactorily explained by DHET at ASSAf's National Scholarly Editors' Forums. That is to say, these indexes are exempted from the DHET's technical evaluative process but the journals indexed are conferred 'accreditation' status.

Universities confer economic value to these lists, firstly because publishing in accredited titles earns variable subsidy from DHET for the authors' institutions in the form of a publication incentive of about ZAR120 000 (the amount varies annually, depending on the DHET annual budget). Secondly, the system provides proof to government of due productivity in the lists that it recognises (rather than also those *additional lists* that academics themselves recognise).

Value chain contradictions

A value chain describes a process view of organisations as a system, composed of subsystems, each with inputs, transformation processes and outputs. An efficiently operating value chain reduces cost, optimises efficiency, eliminates waste and enables competitive advantage (such as in university rankings). Academia, however, is the only industry that deliberately encourages overproduction, irrespective of markets, outlets or consumers (readers, libraries, or retailers). The research has been done, the writing completed, and the papers submitted. But the majority of submissions are rejected by the top international journals, and of those published, many products (articles) might not find a viable readership and are rarely cited. Overproduction leads to over-burdened editors and reviewers; a waste of resources; fewer citations; and opportunity cost of reduced focus on educational activities. The result is less chance of the average academic producing a few valuable articles rather than many worthless articles over their career (De Jager P 2017, written communication, October 15). Social value is equated by universities to measurable economic value, that is, incentive subsidy, thereby increasing pressure on academics to 'perform' (publish, secure grants, advance fields of study and attract third-stream income).[12]

However, the very same articles would most likely find easier access to journals in the domestic DHET-list and to those open-access journals constituted specifically to service and absorb the oversupply of product/articles.[2] The DHET reward for publication is the same across all accredited publications; the preparation, submission and revision costs – in terms of actual time and ability required – are lower and the probability of acceptance is higher for lower-quality journals.[7] In South Africa, this production inflation is met by an oversupply of 'accredited' journals that have emerged to take up the slack. For example, 19 journals service the discipline of management studies in South Africa. An analysis of 371 articles appearing in journals in this field in 2011 revealed a cost to government of subsidising plagiarised work in these journals at ZAR7 million from a total payout to universities of ZAR32 400 000.[13] (See also de Jager et al.[3]).

Overproduction explains the rise of megajournals such as *PLoS ONE* that promise quick acceptance and publication turnaround (Mouton J 2017, written communication, June 6). A knock-on effect is that some South African open-access journals owned by a private company are now engaging in 'market-related' pricing (with regard to the DHET incentive). The example below compares the cost of conventional and open-access publishing. According to the homepage of the open-access journal *HTS Theological Studies*, to use a specific example, the article-processing charge (APC) for publishing in 2017 was ZAR1073 per A4 output page in PDF, with the average article length reportedly 8 pages. Because this journal is indexed in Web of Science, the DHET rule introduced in 2015 pertaining to a maximum of 25% authorship from a single institution does

not apply. That is, an online publication for a single issue can include over a hundred articles when previously 10 might have been feasible in terms of hard copy page allocation. The rule was introduced to discourage rent-seeking by journals that were in the habit of publishing the majority of papers authored by academics from the publication's home base. The previous 50:50 split was recast to 25:75 between 'home' and external authors publishing in any single issue.[14]

Using an issue of *HTS* as an example (issue 73(3) of 2017): nine Unisa-affiliated authors published in this issue, contributing APCs totalling ZAR77 256. In comparison, the typical subscription cost (print and electronic) of two Brill academic theology journals – *Numen* and *Novum Testamentum* – was EUR545 (ZAR8175) and EUR396 (ZAR5940), respectively. Both of these journals are indexed by Web of Science and Scopus, with all the associated benefits of high visibility, ranking and metrics…and with no APCs. Thus, if all nine Unisa-affiliated authors had published in one of these subscription journals, the same access to the same articles would have cost ZAR14 115, and the authors would have received the benefit of a bigger proportion of the subsidy because both these journals are indexed by Web of Science and Scopus. The question is, of course, whether any of the articles would have been published in *Numen* and *Novum Testamentum* for which selection standards might have been more competitive. The comparison suggests that the open-access author-pays model as currently operating in South Africa might not actually be a cost saver. As van den Heever, who provided the above example, concludes:

> *…for the price of publishing 9 articles in one journal, the university could have obtained access to about 12 other journals, a consideration of which, in a context of austerity and cut-backs to library investments and subscriptions, is a factor* [of] *very great importance* (van den Heever G 2017, written communication, September 4).

This kind of pricing nuance between open-access and subscription journals needs to be properly assessed when budgeting research awards and costs. Now overlaid on the budgeting consideration is that performance management regimes require minimum publication targets, in accredited titles, for staff. This 'virtual' dependency has created an institutional equilibrium in which an increasing proportion of South African based academics are wilful (or reluctant) participants in perverse incentive-seeking conduct. It then becomes impossible to separate institutional characteristics from individual behaviour.[7]

Cash cows: Accreditation (the branding) and journals (the factory)

Let us metaphorically compare the practice of accreditation that results in overproduction to the way that a dairy farmer operates with regard to retail markets. The metaphorical discourse of 'cash cows' dominates discussions within university committees. Third-stream publication income is not discussed within committees in terms of 'rent-seeking' behaviour or perverseness. Incentive-seeking discourse framed by the third-stream category predominates. In other words, university managers and committees talk about maximising income via manipulating the possibility of variable income that is enabled by leveraging the DHET publication incentive mechanism to the hilt, and additionally requiring their staff to raise their own funds from funders, donors, research bodies and so on. This is known as 'third-stream' income – the first being (declining) state subsidy and the second being (since 2017, declining) student fees.

In the analogy, let us assume that the dairy owns 100 cows that are milked twice daily for processed products that will be bought by 50 supermarkets which also purchase from other dairies. The supermarkets in the delivery area can only sell 1000 units a day bought from the dairies, so they do not purchase more than that. The dairies do not produce more than they can sell, unless of course they get government subsidies for quantity rather than for sales and consumption. The government subsidies result in overproduction and the mass destruction of unsold milk products. This also keeps the price at a viable level for both farmer and supermarket at the taxpayer's expense.

Likewise, the South African academy insists on overproduction and restricts submission to 'accredited' supermarkets (journals) which are, by default, conferred their qualifying brands by DHET. These 'brands' (lists) then enable the seeking that funds universities but rarely journals. The value chain thus ignores hidden costs, including that of peer review, editing and production. These are delivered 'free of charge' by academics and publishers, but utilise university time and infrastructure – underwritten by the taxpayer.

The cows (i.e. academics), however, get punished for the overproduction of articles placed in non-accredited titles – a glut that is caused by management (the dairy factory) in the first place. Punishment for publishing in journals outside the approved lists takes different forms at different universities. In the weak case, authors are not rewarded and/or their unaccredited publications are not listed in their university's annual report. In stronger cases, such authors are called in by deans for 'counselling', and in the strongest case they may forfeit notch increases, promotion and superannuation. On the other hand, the overproduction could be intellectually valorised by being published in other kinds of fora, including non-accredited peer-reviewed publications, informed and professional magazines, textbooks, subvented books, blogs, letters, commentaries, and so on. However, these outputs are institutionally discouraged because they do not earn DHET subsidy and the social value of such publications is depreciated in research and performance management committees. Yet it was never DHET's intention that publication in the wider unaccredited circuits be discouraged. This consequence has been one of the negative externalities of the way that most South African universities have distorted the policy.

Most universities reward the individual cows (authors) rather than the supermarkets (journals) that take the risk, do the editorial work and publish their output. Many universities top slice a portion of the DHET research incentive of ZAR120 000 for the author's university research code, which can be used for various expenses such as conference fees, employment of student researchers, page fees, book purchasing and article processing charges. Some universities even permit the authors employed by them to appropriate a portion of the incentive as taxable income, thus encouraging active rent-seeking by individual authors, which could be as high as ZAR80 000 an article. The cows thus behave as rent-seekers and keep producing more and more milk (articles) than can be stocked and sold by the approved (mostly non-subsidised) supermarkets (journals).

The expectation of the factories (the universities), as articulated by submitting authors, is that the supermarkets (journals) *must* 'buy' the milk (articles) no matter their capacity. Some journals do offer the cream, for example when some universities reward authors at a considerably higher level for a publication in a high-impact Web of Science indexed journal. But one of the main problems is that overproduction coincides with a dramatic fall in quality (Muller S 2017, written communication, September 4). And, as Phillip de Jager critically observes of the metaphor, '…milk is milk is milk – it is an obvious commodity. Research on the other hand does not need to be a commodity; it can be very valuable and insightful' (de Jager P 2017, written communication, October 15).

There will always be wastage in the system (such as work in progress, incomplete articles, work under revision, unplaced papers), that is, the work that would be more usefully produced as magazine articles or short commentaries. So the debt is sold on down the value chain – to the journals – which do not benefit from incentives, sales or wastage.

The journals – whether funded or not – are, in fact, massively subsidising both the authors and their employers as the publishing costs are rarely recovered by the individual, independent, journals.

What about the workers?

The issue of payment of editors and reviewers is more vexed. Most humanities and social science publications are produced on a shoe-string budget by usually unpaid volunteers (editors, peer reviewers, copy

editors and assistants) while most authors are in paid employment doing directly rewarded work, using their employers' time and infrastructure to conduct research and write their articles. But editors of self-funded journals often undertake the copy editing, layout and design, marketing etc. on their own. A notable exception is Unisa Press – the only university-based journal publisher in South Africa, which publishes 45 journals, of which 8 are not accredited. Some Unisa Press journals that are indexed by Web of Science, Scopus and IBSS are published in cooperation with Taylor & Francis, a commercial publisher, whose total South African portfolio is 72 journals. Unisa Press owns 23 and co-publishes 22 titles. AOSIS – a local, commercial open-access publisher – publishes 37 titles. The National Inquiry Services Centre (NISC) that cooperates with Unisa Press and Taylor & Francis, publishes 30 journals, mainly on behalf of scholarly societies in South Africa. The total number of registered South African journals in all disciplines is about 318.

My analysis now shifts to a single case study of the value chain as an exemplar of how individual journals are subsidising the national research economy. Taking, for example, the 2014 volume of *Critical Arts*: 30 South African based authors published in six numbers earned for their respective institutions ZAR3.6 million. But not a single cent was directly funnelled by either DHET or the universities represented to the journal which enabled the authors' institutions to cash in to this extent. A managing editor was appointed on a 5-year contract when I as the Journal's Editor-in-Chief moved from the University of KwaZulu-Natal to the University of Johannesburg in February 2015. The University of Johannesburg is now subsidising through this post other institutions, but at UKZN, the managing editor position was largely funded through my own DHET subsidies, derived from my own publication of articles in *Critical Arts* and other accredited journals, complemented by fundraising and, to a small extent, page charges. Thus, I personally as the Editor-in-Chief was facilitating huge cash flows to the South African universities to which the Journal's authors were affiliated.

Critical Arts was from 2005 licensed to Unisa Press for the print and Africa market, and to Taylor & Francis for the electronic platform and global rights. Thus, the goose that lays the golden egg, the journal (and its publishers), is accorded a notional value only by DHET. Although some journals levy page (and other) charges, the administrative cost of recovering them is very high and rarely successful. Pieter Rall, Journals Managing Editor at Unisa Press, suggests that APCs are a fair way to recoup some of the expenses incurred. Practically, this would only be an option for universities that benefit from the DHET incentive. Articles for which APCs have been paid could then be open access (perhaps only for South African institutions). Page budgets and subscription rates would have to be adjusted if there was enough uptake (Rall P 2017, written communication, September 5).

Rent-seeking practices include author pressure on editors to leapfrog production schedules. Threats of withdrawal (and actual withdrawal) occur even after the journal has significantly invested in administration and the peer-review process. A DHET representative at the 2017 Future of Publishing conference convened by CREST, ASSAf and other organisations, mentioned instances of abusive telephone calls made to DHET staff regarding articles published in IBSS-indexed journals which were deemed by DHET to be of a predatory nature (based on Beall's List and work done by Mouton and Valentine[4]). Such authors are seeking the rewards attached to measurement systems, rather than the incentive of impacting the discipline, while also mistreating the people administering the system.

The unpaid costs of peer review globally were estimated annually at GBP1.9 billion for 2 million published articles in 2012, from the many millions of submissions[15] to the core international journal population. Reviewers have to be actively recruited and reminded. This voluntary labour associated with peer review is indirectly costed against reviewers' salaries if employed, their pensions if retired, and their savings if under- or unemployed. All the while South African researchers are being invited to offer reviews by predatory journals, which is opportunity cost lost to the country. 'A solution would be for university managers to measure where reviewer effort is going in their universities' (de Jager P 2017, written communication, October 15).

Universities' performance management forms do not always credit editing or reviewing, even as part of official community engagement criteria. In a typical performance management contract, the act of reviewing articles disappears as a minuscule item under 'academic citizenship' that typically counts for 5–10% of one's key performance areas (KPAs). By far the greatest weighting of KPAs at Unisa, as one illustration, is on research, at 30–50%, next to teaching at 30–50%. In the research KPA, the compulsory weighting is 80% for published articles ('bonus weights' of 10% are given for a NRF rating and applying for external grants). The research KPA is the single biggest determinant of an Unisa academic's performance score through which annual bonuses are calculated. The implication is clear: according to university managements, somehow the publication of research just 'takes place', with only the end products valorised. The process of getting research published is completely ignored, or regarded as of little value. Reviewing could be added to one's 'worksheet' on the performance contract. But the KPA that really counts – published articles in 'accredited journals' – is a formalised item to which one cannot add or subtract. Thus reviewing and editing is designed out of the value chain (van den Heever G 2016 written communication, August 31). Such labour becomes an 'after-hours' 'leisure' activity, for which reviewing a single article can take anything from 1 hour to 16 hours. The time:expertise earning ratio is 'written off' as 'service', a donation or unrecoverable expense. The taxpayer pays multiply: (1) taxpayers sponsor the work done by academics, whose universities then, (2) pay to access it in publication form, and who also (3) underwrite the DHET publication incentive. (4) The academic employers subsidise the cost of peer review through their salaries and, finally, (5) editors working during their 'leisure time' subsidise all the components constituting the value chain.

Publication criterion in graduate examination

Another activity that may be seen as 'milking' the DHET publication subsidy, is the requirement of some universities for the conferral of a graduate degree of either: securing acceptance or publication of an article based on the thesis/dissertation in an accredited journal, or proof of submission thereof. A number of legal and ethical issues may arise from this practice:

- Editors, peer reviewers and publishers may be unaware that they are being implicated in concealed but formal examination processes. Thus are they deceived into offering their unpaid labour and expertise to assist an assessment procedure to which they had not consented nor were contracted to undertake.
- A few journals appear to specifically leverage this sector of the cash cow industry, their editors ensuring that they and their own students publish the majority of articles in these in-house but accredited journals.
- Students are overwhelming accredited journals with submissions, often with no guidance on how to write for journal publication. They are thus stretching journals' costs and capacity to process submissions, and are becoming more and more demanding to be included as rent-seekers. Unisa Press's *Communicatio* has responded by specifically excluding submissions based on these criteria – unless new knowledge is conveyed – in its guide for authors.
- Where a journal might accept a publication, examiners might have failed the thesis, or vice versa. Thus could arise legal implications for the journals and universities involved.

Few universities agree to even minimal page charges should such student articles be accepted (unbeknownst as part of the examination process); they want their cake and they want to eat it, and thus *it is the voluntary labour (editors, reviewers, assistants) who pay the price, who subsidise their peers, who cede their intellectual property in their reports to these authors who want someone else to pay for the costs incurred.* ASSAf comments: 'It is recommended that the journal states the submission charges, page fees and APCs in a transparent and simple way, without misleading potential authors' (National Scholarly Editors' Forum circular, 2016 August 29). *The funds for such charges are incorporated in the DHET budget, but they are not allocated appropriately throughout the*

value chain by those who are managing it. In other words, (1) university policies decline or limit the payment of charges; (2) authors who will benefit from the DHET incentive often refuse to dip into this payment or their own funds to cover the charges themselves; and (3) DHET pays the journals nothing at all.

It is a wonderful cow's life for the authors and their universities – they get to eat from the state-sponsored trough. But for alienated editors and self-funded journals, well, they are the ones supplying the trough and the largely free feed.

ASSAf and auditing of journals

This section partially draws on a response invited by ASSAf in 2016 from the Academic and Non-Fiction Authors' Association of South Africa (ANFASA), an advocacy organisation for the protection and advancement of authors' rights.

While university administrations often see journals as cash cows, ASSAf takes a much more critically engaged stance. Its 5-yearly audit cycles attempt to make journals, the missing link in the value chain, but not authors or universities, accountable to the public purse. ASSAf's journals evaluation aims to improve the functioning of the accreditation system, as well as to helpfully encourage the quality of scholarly publication.[16] Regular assessments examine whether or not journals are of 'sufficiently high quality' and meet 'international standards'. The reviewers evaluate (1) scope and focus; (2) editorial and review processes; (3) authorship; (4) enrichment features; (5) financial sustainability; and (6) international positioning. The 'opportunity for corrective action', or suggestions for improvement, is amongst the recommendations offered by ASSAf panels.[17]

However, the task of rectifying the journal subsidy system is extremely difficult as the DHET list has been unevenly evaluated for the 300+ DHET registered journals, and is driven by a greater focus on operations than editorial philosophy, or even quality. ASSAf's assumptions include:

1. The continuing dominance of journal articles as the primary output for research done at South African universities. DHET did however recognise the importance of books in 2015, and also creative work.[14] The new provisions are of special interest to film, television, theatre, dance, video, design, art and fiction writers (and 'plant breeders'). Creative outputs qualify for DHET incentive funding under specific conditions. The legislation does not mention radio, cartooning, motion books (but might include animation as 'film'), digital media or journalism. It does not include non-formal creative interventions, like participatory, forum, educational or street theatre – unless these are research-based and supported by scientific companion outputs also. The legislation specifically excludes self-publishing, but does not define this category. But it does refer to the requirement of a 'credible' publisher able to produce evidence that the work underwent a refereeing process' where novels, poetry, novellas and plays are concerned. This development has resulted from many years of discussion in various fora (NRF ratings committees especially), but the legislation uses archaic (and therefore measurable) categories. 'Authors', like the existing legislation, are assumed to be employees of their institutions, which applies for the award that must undergo a specified peer review process. Peer review significantly includes 'the public domain' – as in theatres, museums and galleries.
2. The continuing significance and effectiveness of accreditation, although DHET is increasingly alert to anomalies.[6]
3. The form of the scholarly journal is assumed as somewhat static, although it is constantly evolving, while the *South African Journal of Science* is offered as the exemplar.
4. Open access is considered by ASSAf[16] to be an unmitigated good, especially in the context of using the evaluation process to identify journals to be added to the state-supported SciELO (Scientific Electronic Library Online) open-access platform. SciELO South Africa is a free-to-access searchable collection of selected, South African, open-access, scholarly journals; inclusion in the platform is free for journals. The project is inspired by a global movement towards the implementation of open-access journals, pioneered by the SciELO project, based in Brazil. In this case it is the platform that is supported, not the journals hosted on the platform.[18]
5. While financial sustainability is a significant aspect for the viability of academic journals, it bears little relation to quality. *If the intention is to improve financial sustainability, then a DHET journals subvention would be of more assistance than a vague assessment of business processes*.
6. International positioning appears to be assessed either by a subjective account of a journal's reputation and reach, or by its listing on the qualifying indexes. However, if such international indexing is so highly valued, then there should be no need to even review journals that are already listed on these platforms – a criterion now accepted by ASSAf.

While ASSAf is aware of many contradictions affecting the DHET system[16,19], some points still need to be flagged. Firstly, a confusion of content with form is evident in the ASSAf questionnaire that asks about hard copy subscription numbers, rather than about subscription bundles. Secondly, some journals that were lauded in some reports have been shown by other ASSAf-commissioned studies presented by CREST at ASSAf-organised National Scholarly Editors' Forum meetings to be overly reliant on authorship from a single institution, in some cases, single individuals seemingly operating as incentive-seeking cash cows. Thirdly, editors are simply assumed to be postmasters shuffling submissions around, when in fact they might be themselves actively shaping a discipline. As one editor of an international journal put it:

> *Shaping a discipline means pursuing the kind of content you judge to encapsulate the vision of the journal – from advertising themed issues, to publishing seminar/conference proceedings, to 'trawling' conference programme books for papers that fit the focus of the journal – apart from rigorously selecting from submissions those papers that add to the ongoing formation of the journal focus* (van den Heever G 2016, written communication, August 3).

The related assumption that competitiveness (in editor and board composition) is better than long-term stability offered by the longer-serving editors and boards is another indication that journals are considered merely as supermarkets from which browsers can choose the products that best suit their own needs. In the humanities, single journals can shape entire disciplines:

> *Long service editors do happen to have deep experience of discipline, methodology and theory, and especially in the case of a journal like* Religion & Theology *that traverses disciplinary boundaries, an editor who has been in the seat for a length of time does have a workable idea of what goes on in a far wider set of disciplinary fields than his or her own specialisation. Innovation in science comes from this kind of 'transcendent' view. It also aids institutional memory* (van den Heever G 2016, written communication, August 31).

The implicit common assumption that all journals are sustainably funded, with fully resourced secretariats, is to be cautioned. Again, van den Heever paints his own – quite common – humanities experience with Brill, a prominent academic publisher based in Belgium:

> *In some cases disciplinary societies pay for editorial help. I have none. I am the proverbial chef, cook, and bottle washer. I handle the Editorial Manager submission platform, I attend to all copy editing, I do proof reading, I sometimes write reviews, review articles, and some headline articles when I deem it necessary to give the lead in the kind of research and articles the journal should publish. All this without much recognition by the university.*

> *Yet, without journal editors there* [would] *not be any published research and no research output subsidies. In our case it is as if the university is of the opinion that somehow publications are 'just there'.*

In contrast, when *Journal for the American Academy of Religion* advertised the position of editor, the advert explicitly stated that, along with all the necessary documentation relating to the applicants' own academic profiles, applicants also had to submit an undertaking from their universities regarding (1) a teaching release; (2) provision of an editorial office; and (3) provision of an editorial assistant. 'This is what it means to take editing journals seriously and doing it professionally', concludes van den Heever (2016 August 31). In terms of rent-seeking, thus one institution effectively subsidises all those who publish in the journal – and often the publisher also.

Conclusion: What is the big deal?

Unless the journals themselves as the most crucial link in the research value chain – and not just universities – are to be funded, sections of the edifice will remain precarious and continued rent-seeking will characterise university research economies, performance management criteria and higher degree administration. The cash cow – the journal – is overburdened, under-fed and producing often sub-standard milk in the absence of sufficient feed.

Structural solutions are required. These solutions include addressing opportunistic institutional rent-seeking morality that has perversely distorted the DHET publication incentive. Overproduction for the sake of the DHET subsidy and key performance indicators should be discouraged. It is also important to recognise that 'some journals only come into existence because of overproduction' (Muller S 2017, written communication, September 4). *The subsidy is for universities, not individuals*. While individuals within universities will benefit, it cannot be ethically allocated as even taxable take-home pay (Di Parker, Deputy Director-General DHET, CREST conference, 2017 September 27).

The value chain must be assessed and funded in its entirety. Journals are the key link – without journals there are no authors. That is, if the now intensively institutionally embedded DHET system is to continue, credible journals must be directly subsidised and rigorously evaluated by ASSAf on the basis of clear and appropriate replicable methodology across journals. If journal support is implemented then the mechanism must be wary of supporting journals that have only been brought into existence by incentive-seeking. This kind of behaviour can be easily moderated more broadly by capping the number of awards made annually to perhaps 10 articles per author.

ASSAf needs not only to assess individual journals by disciplines, but also entire disciplines in terms of what the 'market' will bear, taking into consideration that markets and quality are not necessarily coincident. The case highlighted by Thomas and De Bruin[13] of the oversupply of management journals and the consequent overproduction of articles evidencing significant plagiarism is a clear case in point. Similarly, does South Africa really need 24 law journals and 25 theology titles, which individually exceed the actual number of universities? Assessment might include in exceptional cases a sampling of peer review reports and editorial correspondence. Many journals just use tick boxes which discourage substantive engagement with submissions and authors, while many do evidence detailed critique over one or more drafts. This approach would not in the normal course of affairs necessarily constitute editorial interference by ASSAf panels, as it is after all tasked to assess public accountability.

Current debate should hopefully lead to a sustained discussion of the system that includes both editors and publishers.

Acknowledgements

Thanks to Sean Muller, Johann Mouton, Arnold de Beer, Gerhardus van den Heever, Pieter Rall, Phillip de Jager and Beth le Roux for their comments on different drafts of this article. I thank John Butler-Adam for his extensive critique of my use of the concept of rent-seeking. I also thank ANFASA for providing a venue in which these discussions have been previously flagged (see Le Roux[20]). Finally, David Nothling composed the graphic. An overview of the 2017 Future of Publishing conference and slides can be accessed here: http://www0.sun.ac.za/scicom/?news=dynamics-scholarly-publishing-global-trends-local-responses

References

1. Larivière V, Haustein S, Mongeon P. The oligopoly of academic publishers in the digital era. PLoS ONE. 2015;10(6), e0127502, 15 pages. https://doi.org/10.1371/journal.pone.0127502
2. Van Noorden R. Open access: The true cost of science publishing. Nature. 2013;495:426–429. Available from: https://www.nature.com/news/open-access-the-true-cost-of-science-publishing-1.12676#auth-1
3. De Jager P, Van der Spuy P, De Kock F. Do not feed the predators. S Afr J Bus Manage. 2016;48(3):35–45. https://doi.org/10.2139/ssrn.2858750
4. Mouton J, Valentine A. The extent of South African authored articles in predatory journals. S Afr J Sci. 2017;113(7/8), Art. #2017-0010, 9 pages. https://doi.org/10.17159/sajs.2017/20170010
5. Kerr A, De Jager P. A description of predatory publishing in South African economics departments [document on the Internet]. c2017 [cited 2018 Feb 02]. Available from: https://2017.essa.org.za/fullpaper/essa_3649.pdf
6. Mouton J. Scholarly publishing in SA: The qualitative imperative. c2017 [cited 2018 Feb 02]. Available from: http://www0.sun.ac.za/scicom/wp-content/uploads/2012/10/Mouton-Scholarly-publishing-in-SA-The-qualitative-imperative.pdf
7. Muller SM. Academics as rent seekers: Distorted incentives in higher education, with reference to the South African case. Int J Educ Develop. 2017;52:58–67. https://doi.org/10.1016/j.ijedudev.2016.11.004
8. Tullock G. The welfare costs of tariffs, monopolies, and theft. Econ Inq. 1986;5(3):224–232. https://doi.org/10.1111/j.1465-7295.1967.tb01923.x
9. Krueger AO. The political economy of the rent-seeking society. Am Econ Rev. 1974;64(3):291–303.
10. Tomaselli KG, Steadman I, Gardner S, Muller J, Tomaselli RE, Bertelsen E, et al. Retrospective. Crit Arts. 1982;2:1.
11. Byrne D. Research in a funding jungle: The South African research accreditation system. Scrutiny2. 1996;1(1/2):1–18.
12. Haivan M. Crises of imagination, crises of power. Capitalism, creativity and the commons. London: Zed Books; 2014.
13. Thomas A, De Bruin GP. Plagiarism in South African management journals. S Afr J Sci. 2015;111(1/2), Art. #2014-0017, 3 pages. https://doi.org/10.17159/sajs.2015/20140017
14. Government Gazette 597, 2015 March 11, 38552, Clause 5.0(c)
15. Longva L, Reierth E, Moksness L, Smedsrød B. Peer reviewing: A private affair between the individual researcher and the publishing houses, or a responsibility of the university? J Electron Publish. 2017;20(1). https://doi.org/10.3998/3336451.0020.103
16. Gevers W, Hammes M, Mati X, Mouton J, Page-Shipp R, Pouris A, editors. Report on a strategic approach to research publishing in South Africa. Pretoria: Academy of Science of South Africa; 2006. https://doi.org/10.17159/assaf/0038
17. ASSAf. Peer review panels [webpage on the Internet]. No date [cited 2018 Feb 02]. Available from: http://www.assaf.org.za/index.php/programmes/scholarly-publishing-programme/peer-review-panels
18. ASSAf. SciELO South Africa. No date [cited 2018 Feb 02]. Available from: http://www.assaf.org.za/index.php/programmes/scholarly-publishing-programme/open-access-scielo-south-africa
19. Vaughan CL. Alternatives to the publication subsidy for research funding. S Afr J Sci. 2008;104(3/4):91–96.
20. Le Roux E. Discrimination in publishing. Crit Arts. 2016;29(6):703–704.

Students' ability to correctly apply differentiation rules to structurally different functions

AUTHORS:
Aneshkumar Maharaj[1]
Mthobisi Ntuli[1]

AFFILIATION:
[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

CORRESPONDENCE TO:
Aneshkumar Maharaj

EMAIL:
Maharaja32@ukzn.ac.za

DATES:
Received: 28 Nov. 2017
Revised: 24 Apr. 2018
Accepted: 24 July 2018
Published: 27 Nov. 2018

KEYWORDS:
calculus; derivatives; diagnostics; online quizzes; student difficulties

HOW TO CITE:
Maharaj A, Ntuli M. Students' ability to correctly apply differentiation rules to structurally different functions. S Afr J Sci. 2018;114(11/12), Art. #5008, 7 pages. https://doi.org/10.17159/sajs.2018/5008

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
Eskom Tertiary Education Support Programme; National Research Foundation (South Africa)

The derivative concept is studied in first-year university mathematics. In this study, we focused on students' ability to correctly apply the rules for derivatives of functions with the different structures that they encounter in their university studies. This was done by investigating the online responses of first-year students at the University of KwaZulu-Natal to online quizzes that contributed to their assessment. Based on this investigation, we then interviewed eight students to gain an insight into the thinking behind their responses. We report on the analysis of students' responses to five items on the online quizzes based on the derivative concept. The categories in which those items were based are: condition for existence of derivative at a point; rules for derivatives of standard functions; application of chain rule to different function structures; the application of multiple rules; and application of derivatives to optimise a function. Our findings indicate that students had difficulty in detecting that multiple rules for derivatives were required to differentiate certain types of functions represented in symbolic form. Furthermore, students had difficulty in finding the derivative of a function when more than one application of the chain rule was required. However, there were students who had the ability to apply the rules for derivatives of functions without difficulty. In particular, most of the students were able to correctly recall the differentiation rules for functions with standard structures $f(x)=x^n$, $h(x)=e^{kx}$ and $y=[g(x)]^n$, $n\neq 0$ and k is a non-zero constant. Students were also able to correctly apply the chain rule to an exponential function with base e, raised to $4x$. The majority of students were able to correctly apply the chain rule together with differentiation rules for logarithmic and exponential (with bases $a>1$) function structures, and function structures that required the application of the product rule together with the chain rule. Most of the students were able to apply derivatives to optimise a function.

Significance:
A significant percentage of students who took online quizzes experienced difficulties with applying multiple differentiation rules in the context of a single function. The difficulties stemmed from their inability to detect from the structure of the function which rules should be applied and also the order in which those relevant rules should be applied.

Introduction

One of the most important concepts in university mathematics is the concept of the derivative. In fact, it is one of the fundamental concepts of calculus. In the South African education system, this concept is introduced to learners during their high school studies in mathematics. According to Maharaj[1], first-year university students should already have the knowledge of the concept of the derivative of a function $f(x)$ or f. They are exposed to the following two interpretations of the derivative $f'(x)$ during their schooling years: (1) the gradient of the tangent to the curve f at any point $(x, f(x))$; and (2) the instantaneous rate of change of f with respect to x. While students are introduced to these concepts as early as Grade 12, many first-year university mathematics students have difficulty with the derivative concept.[1] Earlier studies by Orton[2] and Uygur et al.[3] also found that the derivative is a difficult concept for many students. Maharaj[1] carried out a study which used the APOS (action-process-object schema) theoretical framework to investigate university students' understanding of derivatives and their applications in the context of multiple-choice items. A similar approach was used in this study, but the focus here is on students' ability to correctly apply the rules for finding derivatives of functions that they encounter at university level and the application of those rules to find the derivatives of such functions. According to Stewart[4], rules of differentiation help us to calculate with relative ease the derivatives of polynomials, rational functions, algebraic functions, exponential and algorithmic functions, and trigonometric functions. If students have difficulty with these types of calculations, which are regarded as basic, then they are unlikely to correctly apply the concepts that are related to the derivative. Concepts related to the derivative are, for example, increasing or decreasing functions and the concavity of a function, over different intervals. For example, applying the first derivative test to a function $M(x)$ will result in obtaining the increase and decrease intervals of the function where $M(x)$ increases if $M'(x)>0$ and $M(x)$ decreases if $M'(x)<0$. It is our opinion that the students' success in answering a question on the increase or decrease of a function given in symbolic form depends on their ability to interpret the structure of the given function $M(x)$. For this reason we focused on students' ability to correctly apply the rules for differentiation to functions with different structures.

Research question

Are students able to correctly apply the rules for finding the derivatives of functions which have different structures? To help answer this question the following sub-questions were formulated: Which rule(s) can students apply with a high degree of success? To which function structures can they apply those rule(s) successfully? Which rule(s) gave the students difficulty? To which function structures did the students not apply the/those rule(s) successfully?

Literature review

A number of past studies[2,3,5-8] have focused on students' understanding of the derivative concept and how this understanding could be improved. Some of these studies indicated that the derivative is a difficult concept to

understand for many students.[2,3] In particular, students experienced difficulty when applying the rules of derivatives to composite functions.[5] The concept of derivatives forms an important topic of analysis at university level.[7] Hence Maharaj[8] focused on the development of diagnostic testing items for derivatives of functions. The paper by Maharaj[8] was motivated by the need to help first-year mathematics students to improve their performance. That study formulated sample diagnostic questions that could be used to enable students to detect their strengths and or weaknesses. All the questions elaborated on in the results and discussion section of this paper were based on the sample diagnostics questions for calculus that were proposed in the paper by Maharaj[8]. Those sample questions were modified to true/false or multiple-choice questions (MCQs) that were suitable for an online format. For this paper, we analysed the responses of students to five of those modified questions.

Hähkiöniemi[9] stated that exposing students to different kinds of representations can help improve students understanding of the derivative. Zandieh[6] observed that graphical representation is preferred by students when it comes to tasks and explanations about derivatives. The focus in the present study was on the students' ability to correctly apply the rules for differentiation to differentiate functions with different symbolic structures. In this study, in comparison with previous studies in the literature, we assessed whether students were able to correctly detect from the different symbolic structures of functions the rules that were required to differentiate the functions and then apply those rules correctly.

Tall[10] argues that there is a direct link between visualisation and symbolisation when teaching the derivative concept. In this study, we focused on the basic rules for derivatives of functions represented in symbolic form: for example, the basic forms x^n, e^{kx}, $[g(x)]^n$; the derivative of $[g(x)]^n$ is $n[g(x)]^{n-1}.g'(x)$. We further focused on the ability of students to identify the application of such basic rules in the context of particular functions that are also represented in symbolic form [for example $f(x)=3e^{4x}+(5x-1)^e$]; identifying that $(5x-1)^e$ is similar to the structure $[g(x)]^n$ which requires the application of the chain rule. The focus of the investigation was on the students' ability to correctly apply the rules for finding the derivatives of functions which have different structures. Before finding the derivative of a function represented in symbolic form one needs to study the structure of this symbolic form and then make a decision on the rule(s) for differentiation that need(s) to be used. Studying the structure of the function given in symbolic form involves visualisation in the sense that different aspects of the structure have to be seen and noted. For example, when finding the derivative of the function $h(x)=(x^2-x)^{\frac{e}{2}}(1-3x)^{100}$ a student should first study the structure of the function and then decide which rules apply. This function appears as Question 3 in the results and discussion section. Note that the students should recognise from the structure that they need to apply the product rule and the chain rule. Application of the chain rule here is imbedded in the structures of the power rule; for example, in $(x^2-x)^{\frac{e}{2}}$ and $(1-3x)^{100}$. If the student visualises the structure of $h(x)$, detects and notes these different aspects together with the rules for differentiation that are required to be applied, then he or she should arrive at the following:

$$h'(x)=-300(x^2-x)^{\frac{e}{2}}(1-3x)^{99}+\frac{e}{2}(2x-1)(x^2-x)^{\frac{e}{2}-1}(1-3x)^{100}.$$

For students to be able to apply derivatives with a high degree of success, they need to understand the basics of derivatives; this includes unpacking the structure of functions represented in the symbolic form for which the derivatives are required. Rules for finding derivatives of functions help us to calculate with relative ease the derivatives of functions with different structures.[4] It is our view that if students have a high level of understanding of the rules for finding derivatives which are represented in symbolic form, it could be easier for them to apply these rules with a high degree of success to functions of different structures and to concepts related to derivatives.

Conceptual framework

This study was guided by the literature review and the following principles:

1. There is a conceptual hierarchy in the body of mathematics.[11] This principle informed the formulation of the student expected learning outcomes and the development of sample diagnostic questions proposed by Maharaj[11].
2. Students' responses to the items on finding the derivatives of functions given in symbolic form give an insight into their understanding of the rules for derivatives and their applications to finding derivatives of such functions.
3. The quantitative data collected from the relevant online quizzes which focused on students' responses to finding the derivatives of functions represented in symbolic form would reveal trends that could be used to inform teaching with the aim of improving students' understanding of the rules for finding the derivatives of functions.

Methodology and participants

For the 'Introduction to Calculus' module at the University of KwaZulu-Natal, online diagnostics were set up based on the rationale and sample problems outlined in the paper by Maharaj[8]. The problems indicated there were transformed to the format of true/false statements or MCQs that were suitable for online quizzes. These quizzes were a subset of quizzes students were required to take online that contributed to the calculation of their class marks for the module.

A total of 293 first-year undergraduate students were registered for the Introduction to Calculus module at the University of KwaZulu-Natal in 2017. This module is compulsory for those wanting to pursue studies in mathematics. As part of the module assessment, students must undertake online quizzes which assess the material covered in class. The online quizzes were designed in such a way that students could do them anywhere and at anytime within the time frame set for each quiz by the module coordinator or lecturer. The University has numerous computer labs, some of which are open 24 hours a day, so the students had access to computers to take each quiz. Students could also access the quizzes using their own devices, even from outside university premises. The online quizzes were administered by the Moodle platform that was used at the University of KwaZulu-Natal.

The online system provided instantaneous feedback from which the students could determine their strengths and weaknesses when answering a particular question type. The system allowed a student a maximum of five attempts per MCQ but there were penalties for multiple attempts. Each question had a maximum of 2 marks. If a student submitted the correct answer on their first attempt, they were awarded 2 marks; however, only 1 mark was awarded for a correct answer on the second or third attempts. A correct answer on the fourth or fifth attempts scored zero. An incorrect answer on the fifth attempt also scored zero.

The rationale was that, after the students were exposed to formal lectures on a particular section, by taking these online quizzes they could determine their strengths and weaknesses on a topic before sitting for formal written tests. If weaknesses were determined, students were expected to take appropriate remedial actions, for example, revise a section or seek help from a hot seat tutor who was available for individual student consultations at specified times each weekday. The term 'hot seat tutor' refers to a tutor who is available to assist students for particular first-year modules, outside the designated tutorial times. Students who require assistance can access the hot seat tutors during the specified times and meet with them on a one-to-one basis.

At the end of the first semester in 2017, the data for those quizzes were obtained from the Moodle site and the statistics obtained were used in the analysis.

In this paper we focus on five quizzes which covered the section on rules for derivatives and their applications. For this study, only those students who completed all the quiz questions were regarded as having taken the quiz. Students who submitted their responses to only some of the questions were not considered. In the results and discussion section, the focus is on only five quiz items selected from those five quizzes.

Those five items were chosen because they give an overall insight into student responses. For each of those five items, statistics were retrieved from the Moodle site on the following: facility index; discrimination index; and discriminative efficiency. The meanings of these are briefly outlined below as they were used in the results and discussion section for the structure analysis of each of the five quiz items that were chosen. We also give the meanings of intended weight and effective weight, as the latter is used in the description of discrimination index. The reader is referred to https://docs.moodle.org/dev/Quiz_statistics_calculations for further clarity on these terms.

Facility index: Obtained from the mean score of students for an item. The mean score over 2 is expressed as a percentage; the higher the facility index, the easier the question. For a true or false type question the facility index was calculated by using the students' first attempt, and for MCQs, all attempts were used in the calculation. Interpretation of the results is given in Table 1.

Intended weight (IW): The question weight expressed as a percentage of the overall quiz score. Because each item had a maximum score of 2, $IW = 100\,\frac{2}{2\ (\textit{number of items in the quiz})}$. So, for a quiz with four items, the intended weight for each item is 25%.

Effective weight: An estimate of the weight the question actually has in contributing to the overall spread of scores for a given quiz. The effective weights should add to 100%. Note that in the results and discussion section, the five quiz items selected were extracted from five different quizzes.

Discrimination index: The correlation between the effective weight of an item from a quiz and the rest of the items in the quiz, expressed as a percentage. It indicates how effective the item is at sorting out able students from those who are less able. The results were interpreted as indicated in Table 2.

Discrimination efficiency: This statistic is expressed as a percentage of attempts to estimate how good the discrimination index is relative to the difficulty of the question. An item which is very easy or very difficult cannot be used to discriminate students' because most students are likely to get the same score for that item. Maximum discrimination requires a facility index in the range 30–70% (although such a value is no guarantee of a high discrimination index). The discrimination efficiency will very rarely approach 100%, but values in excess of 50% should be achievable. Lower values indicate that the question is not nearly as effective at discriminating between students of different ability as it might be and therefore is not a particularly good question.

Table 1: Interpretation of the facility index based on the students' mean percentage score for an item

Facility index	Interpretation
5 ≤	Extremely difficult or something wrong with the question
6–10	Very difficult
11–20	Difficult
20–34	Moderately difficult
35–64	About right for the average student
66–80	Fairly easy
81–89	Easy
90–94	Very easy
95–100	Extremely easy

Table 2: Interpretation of the discrimination index

Discrimination index	Interpretation
Negative	Question probably invalid
20–29	Weak discrimination
30–50	Adequate discrimination
50 and above	Very good discrimination

After analyses of the data, we emailed 14 students who were selected based on their attempts to correctly answer the five items. The selected students did not submit the correct response on their first attempt. Those who did not submit a correct response even after five attempts were also included in the selection to gain insight into why they were unable to answer correctly. After repeated requests for interviews via email and at tutorial sessions, eight students agreed to be interviewed. During the interview, we accessed that student's online record of submissions to determine which items they answered incorrectly. The student was given a printed copy of the five quiz items. For some items we indicated the student's response. In such cases, the students were asked to explain their responses. The student was allowed to do the relevant working on the print copy or to think aloud. Based on the student's verbal and/or written responses, we probed further to get a deeper insight into the student's reasoning.

The student participants completed an online consent form. Ethical clearance for the study was obtained from the Research Office of the University of KwaZulu-Natal (protocol reference HSS/1058/014CA).

Results and discussion

The results are presented under the following sub-headings:

Question 1: condition for existence of derivative at a point

Question 2: rules for derivatives of standard functions

Question 3: application of chain rule to different function structures

Question 4: application of multiple rules

Question 5: application of derivatives to optimise a function

In each case, the relevant question is given, followed by the question structure analysis and the analysis of student responses to that question. The latter includes relevant extracts from the interviews while question structure analysis focuses on the facility index, discrimination index and discrimination efficiency of the relevant question.

Question 1: Condition for existence of derivative at a point

This question focused on the defining condition for the derivative of a function to exist at a specific value in the domain of the function; finding the derivative from first principles. Basically, one needs to use the formal definition of the derivative based on first principles and use algebra to find a general expression for the gradient of the tangent to the curve f at any point $(x, f(x))$. This question reviews the importance of the formal concept definition.[8]

1. State whether the following statement is true or false. The defining condition for the derivative of a function f to exist at $x=a$ in its domain is that $f'(a) = \lim_{h\to\infty} \frac{f(a+h)-f(x)}{h}$ exists.

   Select one:
   ☐ True
   ☐ False

The facility index indicated that Question 1 was fairly easy while the discrimination index suggested that this question was adequate in discriminating able students from those who were less able (Table 3).

Table 3: Question 1 structure analysis ($n = 278$)

Facility index	Discrimination index	Discrimination efficiency
73.74%	34.02%	39.36%

The students demonstrated a high level of understanding of the derivative from first principles. The facility index of 73.74% gives a clear indication that most of the students did not experience difficulty in responding correctly to the question. It should be noted that this is a definition question and students are expected to answer correctly on their first attempt. That 205 students answered correctly on their first attempt (Table 4) implies that about 26% of the students had difficulty with the defining condition for the derivative of a function at a point. Note that the frequency column in Table 4 indicates, as a percentage for each response to the item, the ratio of the total number of attempts for this response over the number of students who submitted an attempt to this item. For example, for Question 1, the number of 'false' responses is 263, over the number of attempts which is 278, which gives a frequency of 94.60%. The same interpretation applies to the frequency column in the tables that follow. For students who responded 'true', it seems that they did not properly observe and detect the salient feature in the expression for $f'(a)$. This assumption seemed to be confirmed during a think-aloud interview with Student S5 who gave the response 'false' during the interview. When asked why, she responded as follows:

S5: If it is $f'(a)$ why did this [pointing to the x in $f(x)$] *not change to a?*

This student observed the given expression and detected that the $f(x)$ within the given expression for the derivative at a point should be $f(a)$. That was the reason for the response 'false' during the interview. In our opinion, looking at an expression does not imply that one observes the salient features of the expression. If one accepts this, then the ability to actually observe is an aspect that could be focused on and developed among students during the teaching process. So the teaching implication here is that students need to be taught how to observe/see features within the structure of expressions and also within equations that define functions. This focus on observing features within the structure of expressions and also within equations that define functions could be done by framing suitable questions of the type given in Question 1, followed by asking for a reason for the response.

We now focus on the different rules for finding the derivatives of functions which have different structures.

Table 4: Analysis of student responses for Question 1 ($n = 278$)

Response	Credit	Attempt 1	Attempt 2	Frequency
False	100%	205	58	94.60%
True	0.00%	73	0	26.26%

Question 2: Rules for derivatives of standard functions

This question was designed to focus on students' understanding of the power rule and chain rule. Table 5 summarises the structure analysis of Question 2 while Table 6 indicates the student responses.

> 2. Consider the following regarding the derivative of the standard functions with structures $f(x)=x^n$, $h(x)=e^{kx}$ and $y=[g(x)]^n$, $n \neq 0$ and k is a non-zero constant. Select the correct option based on the following:
>
> i. $f'(x)=nx^{n-1}$
> ii. $y'=n[g(x)]^{n-1}$
> iii. $h'(x)=ke^{kx}$
>
> Select one:
> ☐ Only i
> ☐ Only i and ii
> ☐ Only ii and iii
> ☐ None of them

The analysis of student responses indicated that students could apply the power rule on the algebraic and exponential functions with high levels of success. This ability was evident by the fact that about 75% of students chose the correct answer on their first attempt. Note that 54 students chose the first option as their correct answer. This suggested that those students failed to visualise that the exponential function $h(x)=e^{kx}$ was not the standard exponential function e^x, so that the derivative is different from the latter standard function. This was confirmed during the interview when Student S1 was asked why he indicated as his response 'only i'. The relevant extract from the interview follows:

S1: I didn't fully grasp the concept of derivative of e^{kx} …..

R: What do you mean by that?

S1: Because in my understanding it stays the same for e^x … there it is different [pointing to the kx in the context of e^{kx} on the sheet].

When asked why he regarded ii as incorrect, the student wrote: $y'=n[g(x)]^{n-1}.g'(x)$. This implies that he was able to detect that the given structure required the application of the chain rule, although he did not see that when finding the derivative of the structure e^{kx}. This implies that it is crucial in the teaching and learning situation to have interactions based on the subtle features of functions represented in symbolic form, in particular when finding the derivative of exponential functions. These interactions should focus on the base and the exponent of the exponential function.

The facility index (82.54%) suggested that the question was easy but the discrimination index (43.89%) and discrimination efficiency (50.13%) indicated that even though the question was easy it was still effective at discriminating between students of different abilities. In the following question we look closely at the composite function $y=[g(x)]^n$ and see how students who chose the third option for Question 2 found difficulty in the context of finding the derivative of composite functions.

Table 5: Question 2 structure analysis ($n = 272$)

Facility index	Discrimination index	Discrimination efficiency
82.54%	43.89%	50.13%

Table 6: Analysis of student responses for Question 2 ($n = 272$)

Response	Credit	Attempt				Frequency
		1	2	3	4	
Only i	0.00%	48	6	0	0	19.85%
Only i and iii	100%	204	41	12	8	97.43%
Only ii and iii	0.00%	15	8	3	0	9.53%
None of them	0.00%	5	6	5	0	5.88%

Question 3: Application of chain rule to different function structures

The students were required to differentiate the function $f(x)=3e^{4x}+(5x\text{-}1)^e$ and their responses are indicated in Table 7. Table 8 summarises the structure analysis for Question 3.

Table 7 indicates that students were able to differentiate the exponential function structure $3e^{4x}$ with ease. This ease can be concluded by the low number of students who chose the third option. It is interesting to note that those students just used the power rule without seemingly understanding that it does not apply to exponential function structures. This was confirmed during the interview with Student S1; for further details see discussion under Question 4.

The visualisation of the composite function structure and the detection that the chain rule was required in this question was the discriminating factor. Students who were able to unpack the function structure

and detect that the application of the chain rule was required were successful in correctly answering. Table 8 indicates that this question had a discrimination index of 46.25% and discrimination efficiency of 50.19%, hence it was a good discriminator. To arrive at the correct answer, students had to apply the chain rule on the composite function structure $(5x-1)^e$. The first likely difficulty was realising that the exponent e was a constant and that the power rule could easily be applied to this composite function structure. These difficulties could be concluded from Table 7 which indicates that 38 students chose the second option and that 24 students chose the third option during their first three attempts. Those options also indicate that the second difficulty was in applying the chain rule to the composite function structure $(5x-1)^e$. Both these difficulties were detected during the interview with Student S5. The student was asked to differentiate $(5x-1)^e$. Relevant extracts from that think-aloud interview follow:

S5: I don't know what to do.

R: What do you see? [pointing to $(5x-1)^e$]

S5: $5x-1$ *in brackets raised to e.*

R: What is e?

S5: e is a number … so I think we should use the chain rule to find the derivative.

[When asked to do it the student successfully found the derivative of $(5x-1)^e$.]

R: What did you learn from this exercise?

S5: See what is given … don't assume.

This once again implies that it is crucial in the teaching and learning context that students are taught how to see what is given in a symbolic representation of a function.

The above suggests that the correct detection and application of the chain rule was the determining factor on whether a student could or could not arrive at the correct answer. If we look at the first and fourth options note that they differ in the 5 outside the bracket. Students could only arrive at the 5 if they had correctly applied the chain rule.

A suggestion follows on how these particular aspects might be better taught to students. For example, one could give as responses those in Table 7 for the derivative of the function $f(x)=3e^{4x}+(5x-1)^e$. The requirement from students could then be to determine why each response is incorrect or correct. The following illustrative question is framed to set up the teaching activity.

Consider the following four student responses for the derivative of the function $f(x)=3e^{4x}+(5x-1)^e$:

$12e^{4x}+5e(5x-1)^{e-1}$

$12e^{4x}+5(5x-1)^e$

$4xe^{4x-1}+e(5x-1)^e$

$12e^{4x}+e(5x-1)^{e-1}$

Required: Determine whether each of the above responses is correct or incorrect. In each case motivate your answer.

This student activity should be followed by a suitable class discussion based on the answers of students to each given response.

Table 7: Analysis of student responses for Question 3 ($n = 261$)

Response	Credit	Attempt				Frequency
		1	2	3	4	
$12e^{4x}+5e(5x-1)^{e-1}$	100%	179	41	13	13	93.82%
$12e^{4x}+5(5x-1)^e$	0.00%	21	11	6	0	14.67%
$4xe^{4x-1}+e(5x-1)^e$	0.00%	8	11	5	0	9.27%
$12e^{4x}+e(5x-1)^{e-1}$	0.00%	53	10	2	0	25.10%

Table 8: Question 3 structure analysis ($n = 261$)

Facility index	Discrimination index	Discrimination efficiency
75.87%	46.25%	50.19%

Question 4: Application of multiple rules

This question focused on the students' ability to apply differentiation techniques based on different rules to different function structures. It also exposed them to the application of the chain rule in the context of various mathematical representations.[8]

4. Consider the functions defined by:

   $f(x)=ln(5x^2+x)$, $g(x)=(3)^{-x}+log(91-x)$ and $h(x)=(x^2-x)^{\frac{e}{2}}(1-3x)^{100}$. Work out the derivatives of the functions f, g and h. Select the correct option based on the following:

   i. $f'(x)=\frac{10x+1}{5x^2+x}$

   ii. $h'(x)=-300(x^2-x)^{\frac{e}{2}}(1-3x)^{99}+\frac{e}{2}(2x-1)(x^2-x)^{\frac{e}{2}-1}(1-3x)^{100}$

   iii. $g'(x)=-x(3)^{-x-1}-\frac{1}{ln10(91-x)}$

   Select one:
   - ☐ Only i
   - ☐ Only i and ii
   - ☐ Only iii
   - ☐ Only i and iii

This question focused on the application of multiple rules for differentiation in the context of the three functions. The analysis in Table 9 indicates that 46 students chose the first option while 47 students chose the fourth option. This implies that the majority of the students who attempted this question later in the semester were comfortable with applying differentiation rules in the context of exponential and logarithmic functions.

Student S1 was one of the students who answered 'only i and iii'. In the context of the function $g(x)=(3)^{-x}+log(91-x)$ we were interested to know how he obtained the derivative of $(3)^{-x}$. The following is an extract from the interview.

S1: Using the power rule … [and writes $x(3)^{-x-1}$].

R: What is the power rule?

[The student wrote $y=x^n$ followed by $y'=nx^{n-1}$. This was followed by drawing his attention to where the variable x was in the structure x^n]:

S1: In the base.

R: Where is the x in the structure $(3)^{-x}$*?*

S1: [pointing to the x in $(3)^{-x}$] *… can't use the power rule … I need to go and learn this …*

This example reinforces the need in the teaching and learning situation to have interactions based on the subtle features of functions represented in symbolic form; in context of the power and exponential functions this should be with regard to where the variable appears. In particular, differentiation of functions with the following structures should be focused on: $(3)^{-x}$; $(-x)^3$.

A more in-depth analysis of the data relating to those students who chose the second option for their second or later attempts (Table 9) revealed that they had difficulty in differentiating the function $h(x)$. A possible reason for this difficulty could be that the function $h(x)$ has a structure which requires the application of the product and chain rules for differentiation, and more than one application of the chain rule. Our interviews with students indicated that any one of these three – product rule, chain rule or more than one application of the chain rule – could be the reason for their difficulties. For example, students S2 and S3 did not see that the product rule had to be used and Student S7 did not detect the need for the chain rule. Students S4, S5 and S6 detected that both the rules had to be applied, but they applied the rules in the incorrect order:

the chain rule was applied first to both functions in the context of $h(x)$. The following is an extract from the think-aloud interview with Student S5, after the student was asked to differentiate $h(x)=(x^2-x)^{\frac{e}{2}}(1-3x)^{100}$:

R: What do you see?

S5: $(x^2-x)^{\frac{e}{2}}$ *and* $(1-3x)^{100}$ *are multiplied.*

R: To differentiate $h(x)$ *which rules would you apply?*

S5: The chain rule and the product rule.

R: Which of these rules would you apply first?

S5: The chain rule.

When asked to apply the rule, the student wrote down $\frac{e}{2}(x^2-x)^{\frac{e}{2}-1}(2x)\times100(1-3x)^{99}$ (3). We note that the application of the chain rule to both the functions that comprise $h(x)$ is incorrect.

The above, as well as the fact that some students answered incorrectly on their fourth attempt (Table 9), indicated that Question 4 was effective at discriminating able students from those who were less able (discrimination index of 41.60% and discrimination efficiency of 45.42%; see Table 10).

Table 9: Analysis of student responses for Question 4 ($n=259$)

Response	Credit	Attempt				Frequency
		1	2	3	4	
Only i	0.00%	38	6	2	0	17.76%
Only i and ii	100%	177	43	12	4	93.44%
Only iii	0.00%	12	6	2	0	7.43%
Only i and iii	0.00%	32	15	0	0	10.15%

Table 10: Question 4 structure analysis ($n = 259$)

Facility index	Discrimination index	Discrimination efficiency
76.64%	41.60%	45.42%

Question 5: Application of derivatives to optimise a function

This question focused on the application of derivatives in the context of optimisation of a function.

> 5. The function $M(x)=-\frac{1}{45}x^2+2x-20;\ 30\leq x\leq 65$; is an approximation for the number of kilometres per litre of fuel used by a new prototype car, when driven at a speed of x kilometres per hour. Choose the correct option based on the following statements:
>
> i. The number of kilometres per litre of fuel used increases on the speed interval (30,45) and decreases on the speed interval (45,65).
>
> ii. The absolute maximum number of kilometres per litre of fuel used is 25 km per litre.
>
> iii. The absolute maximum is achieved at a speed of 45 km per hour.
>
> Select one:
> ☐ Only iii
> ☐ Only ii and iii
> ☐ Only ii
> ☐ All of them

From Table 11, the discrimination index (37.04%) and the discrimination efficiency imply that the question was not a good discriminator in effectively sorting the able students from those less able. In the context of the data that were available, we could not detect which part of the question contributed to the relatively weak discrimination effect.

Table 11: Question 5 structure analysis ($n = 268$)

Facility index	Discrimination index	Discrimination efficiency
73.30%	37.04%	43.48%

Table 12: Analysis of student responses for Question 5 ($n = 268$)

Response	Credit	Attempt					Frequency
		1	2	3	4	5	
Only iii	0.00%	16	16	4	0	1	13.70%
Only i and iii	0.00%	49	11	4	0	0	23.70%
Only ii	0.00%	12	12	7	0	0	11.48%
All of them	100%	191	12	19	13	1	94.81%

Question 5 was based on derivative-related concepts for a function given in symbolic form, for example: determining the interval(s) for which the function is increasing or decreasing; optimising a function. The question required a student to detect the relevant derivative-related concept within each statement and to do the necessary working to determine if the statement was correct or not. Table 12 indicates that 191 of 268 students (about 71%) correctly answered on the first attempt. This finding implies that a large number of the students were able to detect the relevant derivative concept on which each given statement focused, do the necessary working and make relevant conclusions. What was concerning, is the number of attempts required for some students to obtain the correct answer. This conclusion can be drawn by looking at the first three incorrect options; the students made up to five attempts. In our opinion, such students do not have the necessary derivative-related concepts for a function given in symbolic form, to answer a question of this type. This opinion is supported by the following that transpired during the interview with Student S8:

R: How can you use $M(x)$ *to find out where the function is increasing or decreasing?*

S8: Take the derivative of $M(x)$*, equate it to zero, then solve for* x *…….*

[Student S8 then correctly did the working and arrived at $x=45$, which indicated that the student was able to do a routine procedure, by following an algorithm.]

However, the extracts that follow indicate the student did not know what $x=45$ represented, in the context of the relevant derivative-related concept.

R: Look at $x=45$ *in the context of the interval* $30\leq x\leq 65$*. What can you conclude?*

S8: $x=45$ *lies within the given interval, then the function* $M(x)$ *is an increasing function.*

This response clearly suggests that the student could not interpret that $x=45$ represented the value at which the derivative is 0, although this was part of the working that this student correctly did when following the algorithm. This was confirmed by the silence that followed when the researcher posed the following question: $M'(x)$, *what does this represent?* Upon further probing, the student was able to interpret $x=45$ as confirmed by the following interview extract:

R: What type of function is $M(x)$*?*

[Student S8 was able to correctly identify the function and draw a rough sketch.]

R: Where is the turning point?

[Student S8 correctly pointed out the turning point on the sketch.]

R: What is the value of M'(x) at the turning point?

S8: zero.

R: So, what does x=45 represent?

S8: The x value of the turning point.

From the above, one could conclude that even if students correctly apply the rules for differentiation together with relevant related algorithms, they do not necessarily understand the deeper derivative-related concepts. The implication is that the teaching and learning of the derivative-related concepts should focus on understanding why certain steps in an algorithm are followed. This means that before the algorithms are stated there needs to be understanding of why certain steps are included in the algorithm. In our opinion, this reinforces the need in the teaching and learning situation to have interactions based on the subtle features of functions or equations that result from them, represented in symbolic form.

Conclusions and recommendations

This paper was based on sample diagnostic questions for the concept of derivatives with the aim of improving students' ability to correctly apply the rules for finding derivatives of functions. The study has confirmed that the derivative is one of the concepts students have difficulty with, as indicated in the literature.[2,3] More specifically, students experienced difficulties with applying multiple differentiation rules in the context of a single function (a composite function or imbedded function). These functions required application of the chain rule and it was found that, especially when more than one application of this rule was required, in the context where the applications of multiple differentiation rules were required, students experienced difficulties. It seems that the difficulty stemmed from the inability of students to detect from the structure of the function which rules should be applied. In particular, during teaching, we recommend that there should be a deliberate focus on the different rules that are required to differentiate functions with symbolic forms in the context of exponential and power functions. For example, correct interpretation of the symbolic structure and rules that are required to differentiate each of 3^x and x^3. This should be followed by 3^{-x} and $(-x)^3$ when focusing on application of the chain rule. Although this teaching implication was suspected from the analyses of online responses of students, it was confirmed during the interviews with selected students. In cases in which students detected the rules that were required, some had difficulty in detecting that more than one application of the rule was required. Any one of the three rules – product rule, chain rule or more than one application of the chain rule – could be the reason for students experiencing difficulties. In the context of a function that required the application of the product and chain rules, some students only detected the chain rule and did not see that the product rule was required. Further, students who saw that both rules were required tried to first apply the chain rule to both the functions that comprised the given function, $h(x)=(x^2-x)^{\frac{e}{2}}(1-3x)^{100}$. In the teaching and learning situation the implication is that there should be interactions based on the subtle features of functions represented in symbolic form. If this is accepted, then our recommendation is that lecturers should during formal lectures focus on the importance of studying and visualising the structure of a function. In particular, students need to be taught how to observe/see features within the structure of expressions and also within equations that define functions.

It seems that if students could first study and visualise the given structure of the function, then detecting and noting the structural representation could help them to decide which rule(s) to apply when finding the derivative of the relevant function. We recommend that research be conducted to further investigate this hypothesis.

Acknowledgements

We thank the University of KwaZulu-Natal for allowing us to use the data on Moodle and also the students' results for research purposes. M.N. acknowledges A.M. and Dr D. Varghese. We also acknowledge the NRF for making available grants for the project, 'Online diagnostics for undergraduate mathematics'. The Tertiary Education Support Programme of Eskom is acknowledged for making funds available for the UKZN-Eskom Mathematics Project.

Authors' contributions

A.M.: conceptualisation, critically reviewing the writing, and project leadership; M.N.: writing the initial draft.

References

4. Maharaj A. An APOS analysis of natural science students' understanding of derivatives. S Afr J Educ. 2013;33(1), Art. #458, 19 pages. https://doi.org/10.15700/saje.v33n1a458
5. Orton N. Students' understanding of differentiation. Educ Stud Math. 1983;14(3):235–250. https://doi.org/10.1007/BF00410540
6. Uygur T, Özdaş A. Misconceptions and difficulties with the chain rule. In: Rogerson A, editor. The Mathematics Education into the 21st Century project. Proceedings of the Eighth International Conference; 2005 November 25 – December 01; Skudai, Johor, Malaysia. Malaysia: RIC Publications; 2005. p. 209–213.
7. Stewart J. Calculus. 6th ed. Toronto: Thomson Brooks/Cole; 2009.
8. Tall D. Students' difficulties in calculus. In: Proceedings of Working Group 3 on Students' Difficulties in Calculus, ICME-7; 1992 August 17–23; Québec, Canada. Québec: Les Presses de l'Université Laval; 1993. p. 13–28. Available from: http://homepages.warwick.ac.uk/staff/David.Tall/pdfs/dot1993k-calculus-wg3-icme.pdf
9. Zandieh M. A theoretical framework for analysing student understanding of the concept. CBMS Issues Math Ed. 2000;8:103–122. https://doi.org/10.1090/cbmath/008/06PMid:25265997
10. Orhun N. Graphical understanding in mathematics education: Derivative functions and students' difficulties. Procedia Soc Behav Sci. 2012;55:679–684. https://doi.org/10.1016/j.sbspro.2012.09.551
11. Maharaj A. An outline of possible in-course diagnostics for derivatives and integrals of functions. Int J Sci Educ. 2015;11(1):78–90. https://doi.org/10.1080/09751122.2015.11890377
12. Hähkiöniemi M. Perceptual and symbolic representations as a starting point of the acquisition of the derivative. In: Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education. 2004(3):73–80. http://emis.ams.org/proceedings/PME28/RR/RR168_Hahkioniemi.pdf
13. Tall D. A sensible approach to the calculus. Presented at: The National and International Meeting on the Teaching of Calculus; 2010 September 23–25; Puebla, Mexico. Available from: http://homepages.warwick.ac.uk/staff/David.Tall/pdfs/dot2010a-sensible-calculus.pdf
14. Maharaj A, Wagh V. An outline of possible pre-course diagnostics. S Afr J Sci. 2014;110(7/8), Art. #2013-0244, 7 pages. https://doi.org/10.1590/sajs.2014/20130244

The appropriateness of a realist review for evaluating the South African Housing Subsidy Programme

AUTHORS:
Matodzi M. Amisi[1]
Lochner Marais[2]
Jan S. Cloete[2]

AFFILIATIONS:
[1]South African Department of Performance Monitoring and Evaluation, Pretoria, South Africa
[2]Centre for Development Support, University of the Free State, Bloemfontein, South Africa

CORRESPONDENCE TO:
Lochner Marais

EMAIL:
MaraisJGL@ufs.ac.za

DATES:
Received: 04 Sep. 2017
Revised: 16 Feb. 2018
Accepted: 13 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
critical realism; housing assets; housing policy; poverty

HOW TO CITE:
Amisi MM, Marais L, Cloete JS. The appropriateness of a realist review for evaluating the South African Housing Subsidy Programme. S Afr J Sci. 2018;114(11/12), Art. #4472, 9 pages. https://doi.org/10.17159/sajs.2018/4472

ARTICLE INCLUDES:
✓ Supplementary material
× Data set

FUNDING:
None

Conducting meta-reviews of government programmes has become common practice. In South Africa, the national Department of Human Settlements and the national Department of Performance Monitoring and Evaluation recently commissioned a team to review the extent to which the Housing Subsidy Programme had provided assets to municipalities and the poor and whether these assets had helped poor households escape from poverty. A realist approach was employed to conduct the review. We argue that, given the complex nature of housing programmes, the realist review methodology was an appropriate approach to follow in answering the review questions. We explored how the realist review method allowed us to work with the uneven and contested nature of the housing literature and how the review nonetheless enabled elucidation of the factors that had contributed to the expected outcomes. Because this case was the first time that this method was used in a government-commissioned evaluation of housing, there were some practical challenges involved in its use. Some of the challenges were related to the nature of the questions that were asked. At the time of the review, the Department of Human Settlements was in the process of reviewing the 1996 White Paper and, to inform this process, the Housing Subsidy Programme review included a copious number of questions set by the Department of Human Settlements and Department of Performance Monitoring and Evaluation, which made the review rather large and, in some cases, complicated the analysis. In some cases, because the Departments wanted clear-cut answers, the commissioners perceived the theoretical strength of the method, such as offering explanatory instead of conclusive judgement, as a weakness. The paper reveals some limitations of the realist review method for evaluating the multifaceted outcomes of a complex programme, particularly the practical difficulty of dealing with large quantities of data. We do however consider this method to have potential for further reviews.

Significance:

- Housing research in South Africa is uneven which makes any review process difficult.
- The review was unable to offer judgement on the effect that the Housing Subsidy Programme has had on the asset base of the poor.
- The review was useful for making clear which factors will help the Programme to achieve the intended outcomes and also for pointing out on what government should focus to build assets for the urban poor.

Introduction

Evaluation and review of policy has become a common government practice across the globe. Many of these reviews take the form of meta-reviews, in effect studies of studies, in which the literature pertaining to specific policy concerns is closely examined. The demand for policy reviews has spawned an array of review methods: systematic, realist, scoping, critical, mapping – to mention but a few. In this paper, we assess the 'realist review' method, originated by Pawson and Tilley[1]. For simplicity we have chosen to use the term 'realist review', while noting that this method is also referred to as 'critical realist review', as it stems from critical realist philosophy.

Globally, there is a growing body of work of evaluation of conventional review methods such as systematic reviews.[2] Some common criticisms are that the available evidence is often 'mixed or conflicting' and provides 'little or no clue as to why the intervention worked or did not work when applied in different contexts'[3], that there are difficulties in striking a balance between rigour and relevance, and that 'few review types possess prescribed explicit methodologies and many fall short of being mutually exclusive'[2]. Substantially more work is needed to evaluate review methods[4], particularly in the health sciences[5].

In South Africa, as elsewhere, evaluations and policy reviews have become the norm now that policy is increasingly expected to be evidence based.[6] The national Department of Performance Monitoring and Evaluation (DPME), established in the Office of the President in 2010, has been mainstreaming reviews of policies and programmes in various line departments.

By the end of 2016, the DPME had completed 65 evaluations, 2 of which were meta-reviews. In 2014, the national Department of Human Settlements (DHS) and the DPME commissioned a review of South Africa's Housing Subsidy Programme. The review was to investigate the extent to which the Programme had succeeded in providing assets to the poor and whether these assets had helped poor households escape from poverty. We initially suggested the use of systematic review methodology to answer the review questions. Discussions with both the DPME and the DHS alerted us to the limitations of the systematic review approach in regard to the Housing Subsidy Programme. Because this Programme is implemented non-uniformly by the nine provinces – with each province using different implementation protocols in response to particular local contexts and moreover doing so in a

variety of communities – our main problem was that we needed a review methodology that would be more flexible and would emphasise different contexts. In the end, we opted for the realist review method.

Apartheid planning left South African cities not only with large numbers of informal settlements and housing backlogs but also with municipalities that were ill prepared to accommodate rapid growth. The Housing White Paper released in 1995 was one of the first post-apartheid policy responses to the housing challenges faced by South African communities. Although multifaceted, the policy chiefly emphasised three things: ownership, a focus on the poor (only households with incomes of below ZAR3500 per month are able to access the subsidy) and a fixed-amount capital subsidy. (In 1995, the USD:ZAR exchange rate was 1:3.61 and about 1:13 at the time of writing in July 2017.) The original capital subsidy amount in 1995 was ZAR15 000 for those households with the lowest incomes. A revised policy, namely 'Breaking New Ground: A Comprehensive Housing Plan for the Development of Integrated Sustainable Human Settlements'[7], has retained the above three elements while re-emphasising informal settlement upgrading and rental accommodation, and drawing attention to the need to establish sustainable settlements and to develop the property market. The South African Housing Subsidy Programme has delivered approximately four million housing opportunities (subsidised houses and site-and-services) in slightly more than two decades, mostly by providing a capital subsidy and homeownership to households at the lower end of the market.[8]

Despite the growing number of reviews and internal evaluations in South Africa there has been virtually no critical assessment of their methods. Against the above background, we critically assess the method we used and then discuss its appropriateness in terms of evaluating the multifaceted outcomes of the Housing Subsidy Programme. The fact that we as the authors represent both the commissioning department (the first author) and an academic department should ensure a balanced view. While we acknowledge that our closeness to the review process influenced our evaluation of the review, we did attempt to take a step back. We reflected with hindsight on what had helped or hindered the review process and its outcomes. In this paper, we discuss the limitations, and some benefits, of the realist review method.

Realist reviews: An overview

Realism is a school of thought that lies between positivism and constructivism.[9] Pawson and Tilley[1,5] are credited with applying realist philosophy to programme and policy evaluation. The value of the realist method lies in its ability to deal with complexity[3], to synthesise evidence while accepting that 'no deterministic theory can always explain or predict outcomes in every context'[10]. Evidence-based policy development is commonly described as wanting to determine 'what works'. However, in a realist review, we ask a more complex question: What is it about this programme that works for whom in what circumstances?[3] In a realist review, the reviewers are able to engage with context and the human element in the implementation of interventions. There is an acceptance that different conditions contribute to programme success or failure[1,7,11] and that while diverse results are problematic, various outcomes are inevitable because the mechanisms that create change are not necessarily embedded within a specific programme but are often present in the thought processes of the programme's participants.[1] These diverse results must thus be explored rather than controlled.[12] A realist review therefore emphasises 'what works for whom, in what circumstances, in what respects and how'[3]. Realist evaluators can use both quantitative and qualitative research methods.[3]

Realist reviewers engage with evidence by studying the interaction between contexts, mechanisms and outcomes, in what are called CMO (context-mechanism-outcome) configurations.[8,9] A CMO configuration is 'a proposition stating what it is about an initiative that works', in other words, an hypothesis to be tested.[13]

Conventionally, evaluators find it difficult to deal with how context mediates and moderates the results of a programme. Context is both perceived and treated as a threat to the external validity of evaluation where evaluators are concerned with isolating how programme interventions produced observed outcomes.[14] Realist review methodology, however, allows evaluators to explore a variety of contexts and they try not to be judgemental.[3] Understanding how context mediates and moderates programme performance is thus core to realist reviews. Mechanism is another central component of realist reviews. A realist review looks at the underlying causes of change that are not directly observable.[7] Mechanisms could involve multiple individuals engaged in a sequence of processes.[3] Mechanisms connect programmes to their outcomes. Realist review sees the outcomes as the result of interaction between the resources or opportunities the programme provides, the reasoning of its target population, and the context. The change process is studied to provide explanations for *how* change happens, not just to state what change has been observed.

Other principles besides the CMO configurations underpin a realist review. Firstly, a realist evaluator sees programmes as theories.[1] People design programmes on the basis of their beliefs about the nature of the problem and how change happens. This design is then translated by practitioners who are responsible for delivering services to programme beneficiaries. Thus, programmes are always inserted into existing social systems that have produced the negative conditions that necessitated the programme.[1] Because an intervention may involve multiple theories, using traditional review methods is difficult. In this regard, Pawson et al.[3] note that 'the review question must be carefully articulated so as to prioritize which aspects of which interventions will be examined'. Programme motivations and designs usually make statements about how the programme or policy should be implemented and what results can be expected. Because a realist review usually starts by adopting the programme or policy design as the theoretical base, it must therefore consider the theory's underlying assumptions.

Secondly, as programmes are embedded in social systems, it is 'through the workings of entire systems of social relationships that any changes in behaviours, events and social conditions are effected'[1]. A realist review therefore recognises and accepts the existence and interplay of multiple social systems. To understand the process of change, the reviewer needs to investigate beyond what the programme offers so as to understand how the wider social systems affect the programme. Traditional review methods are often unable to deal with this multiplicity and with interconnections in society. The realist review accepts that the relationship between mechanisms and outcomes does not have to be linear; in many cases it could be a reverse relationship. In accepting the existence of non-linear relationships, the realist reviewer notes and examines the 'flows, blockages and points of contention'[3]. For example, the outcomes in societies that emphasise self-help might prove to be totally different from those in societies in which the state is required to play a dominant role. A second example relates to the fact that while the South African Housing Subsidy Programme grants individual households decision-making status, the decisions that households make might not be all that similar.

Thirdly, programmes are active. Implementation of a programme requires the active participation of individuals.[1,7] This principle is important and has methodological implications. For the realist reviewer, there is no need to control and remove the human influence. Instead, the reviewer needs to explore and understand how the human influence produces change in the intended programme.[1] In a realist review the literature review can therefore be broader than in a traditional review in which control and adherence to predefined programme components, population, types of studies, and so on, are critical. A realist review includes literature on the basis of relevance rather than restricting itself to a pre-identified finite set of sources. It generally uses a simple search strategy based on purposive sampling but multiple search strategies can also be used, and grey literature can be given a more important role than in other review types.

Lastly, because programmes are open systems, realist reviewers accept that externalities will always influence the way in which a programme is implemented, with benefits varying according to location. The programme implementer is an active agent in the implementation of the programme and context will constrain what is implemented.[1,7] Programmes can also be self-transformational. As the programme is

implemented, it may be altered according to lessons learnt and may be adapted to context changes that have resulted from the introduction of the programme. A realist review must therefore be able to account for this adaptability. This aspect was important in the review that is the topic of the present paper, as the Housing Subsidy Programme policy had evolved significantly since 1994. From having an initial focus on starter houses in whose growth households were required to invest, the policy now makes provision for fully built houses of good quality that are aimed at incentivising market take-off.[15]

Realist reviews are not free of limitations. Realist review methods have been criticised for not being able to provide definitive answers to policy issues. The practical applicability of the realist approach has also been called into question, with some arguing that although, theoretically, the method offers useful lenses with which to look at programmes, it is difficult to apply these lenses with the methodological rigour and precision required of evaluators. A widely contested issue is how realist reviewers define and interpret causation. Realist reviews tend to emphasise contextual knowledge (what works for whom in what context) over normative positions; and then, too, the nature of causation is often debatable.[16,17] Effectively, realist reviews should pay attention to how existing world views influence specific studies and researchers' interpretations of the results. The danger further exists that researchers will choose literature that is in line with their own epistemological and ontological assumptions. Further criticism is that there is too little emphasis on the question 'does it work?' (as opposed to what works under what conditions) and an over-emphasis on contextual factors.[13] It is these very criticisms that have necessitated this paper, which reflects on the practical use of the method while attempting to answer a policy question in a complex government programme.

Background to the Programme and implications for the review

South Africa's government-subsidised Housing Subsidy Programme is a complex intervention both in design and mechanisms for implementation (Figure 1). It is complex firstly because it has to respond to dysfunctionalities inherited from the apartheid government. The *Group Areas Act of 1950* moved most black people from the core urban areas to impoverished and marginalised townships. Landownership for black people was revoked during the 1950s and only selectively reinstalled in the second half of the 1980s. The resulting inequality between black and white households should not be underestimated. As a result, the Housing Subsidy Programme was central to the political negotiations during the transition from apartheid to democracy and was important for restorative justice.[18] Housing is now both a constitutional right that the state has an obligation to realise progressively (as affirmed in the Constitutional Court case of the Government of the Republic of South Africa vs Grootboom in 2000) and an individually owned asset that functions in the property market.[15,19,20] Responding to apartheid property-ownership biases (in urban areas), the Housing Subsidy Programme adopted an ownership model designed to redistribute wealth, ensure the participation of the poor (particularly black and coloured people formerly denied ownership in urban areas), and enable households to access and benefit from the workings of the property market.[21,22] The intervention logic or the theory of change was thus always more than the mere provision of accommodation. The provision of accommodation was a means to reduce asset poverty, address the failings of the market, give the poor equitable access to the property market and create wealth for those previously excluded (Figure 1).[23] The 2004 Human Settlements Strategy added to this a clear focus on asset creation as a means of poverty alleviation.[24] The theory of change was thus a market-based approach to asset building. Furthermore, when the Programme started, we had to accept the theory of change because it was the policy position adopted by the DHS. Later in the paper we note that during the review process we started to question this one-dimensional asset-building approach.

A second source of complexity is that the outcomes of the Housing Subsidy Programme are contingent on factors beyond its control or influence. Among these factors are macroeconomic conditions (employment, interest rates, and so on), concomitant investment in public spaces by local government, provision of municipal services, and the socio-economic conditions of the beneficiaries.

Thirdly, the intervention is complex because of its delivery arrangements. Nine provincial Departments of Human Settlements annually deliver housing by means of thousands of construction projects using a range of delivery arrangements with municipalities and private contractors. The nine provinces vary considerably in the way in which they package housing projects, select and appoint building contractors, monitor adherence to policy objectives, work with local governments to secure the spatial planning and other planning approvals necessary for project delivery, and provide bulk services such as water and sanitation. They also vary in the way they plan development so as to integrate low-income households with the rest of the municipality. A further complication is that architects and town planners make decisions about settlement design and land-use schemes (that in turn influence the development trajectory of a settlement). These decisions are made on a project-to-project basis so as to optimise the effective use of land and other resources.

Finally, the households that benefit from government housing subsidies vary in terms of economic circumstances, size and composition, level of education, and so on. To qualify for a subsidy a household must have a combined monthly income of no more than ZAR3500. But households in this income category may be unemployed and dependent on government grants, or formally employed with the possibility of upward economic mobility. They may be single-parent or two-parent households. The type of household determines or influences the extent to which a house will be an asset to that household and how well it will use the resources provided by the Programme. Variation in outcomes is thus only to be expected. Isolated studies on whether housing is elevating people out of poverty are likely to reach different conclusions.

All these complexities had implications for the review. In addition to the ideological context, we had to know the background of papers on housing delivery, such as in which province the research was conducted and the terms of the contractual relationships between developers, contractors and the provincial governments. We also had to take into account the fact that most housing research is currently being done in urban contexts and chiefly in four or five of the largest metropolitan areas, which, although not necessarily a negative, could give our review an urban bias. These factors significantly influence the ability of the Housing Subsidy Programme to achieve its policy objectives and tend to make the delivery mechanisms unduly dependent on context.

The review

The review was commissioned by the DHS and the DPME as part of the cabinet-approved National Evaluation Plan of 2013/2014. The DPME is the custodian of the Plan, as part of the implementation of the National Evaluation System. After 20 years of implementing the Housing Subsidy Programme, DHS reviewed its housing policy to respond to the transition to a broader human settlements approach initiated by the 2004 Breaking New Ground strategy and mandated in 2009 with the name change from 'Department of Housing' to 'Department of Human Settlements'. Our review was one of seven evaluations that the DHS conducted in partnership with the DPME, intended to influence and inform this policy review process. The need for a review emanated from this policy need.

The review questions

The review's specific focus was to 'determine if the provision of state subsidised housing [had] addressed asset poverty for households and created assets for municipalities'. More specifically, the review questioned whether subsidised houses were 'growing in value' and whether beneficiaries were indeed obtaining and benefitting from this growth. A set of 14 secondary questions pertained to the theoretical and conceptual understanding of housing and assets, asset generation for individual households and asset generation for municipalities (see Appendix 1 in the supplementary material for a full list of questions).

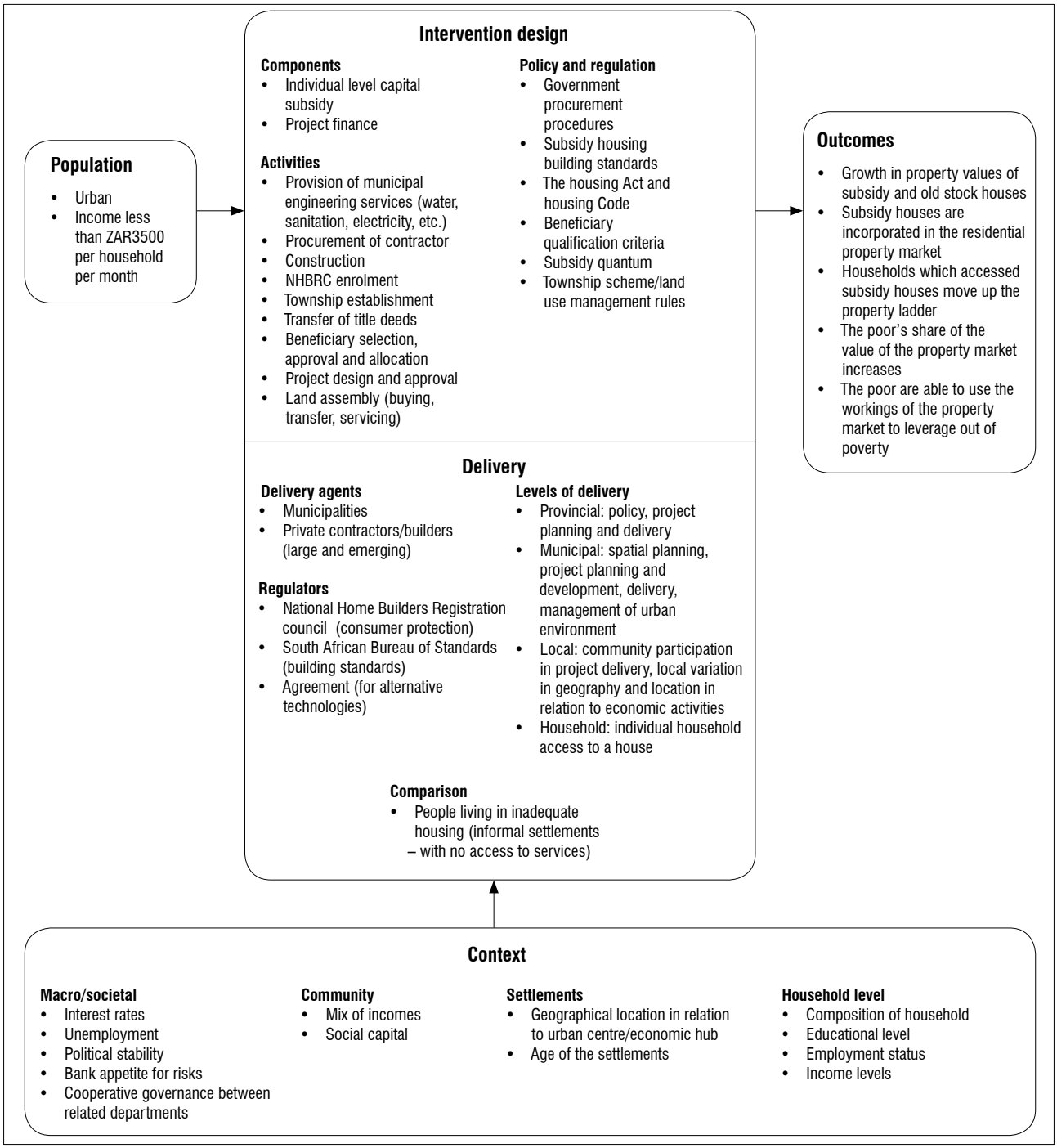


Figure 1: Population-intervention-comparison-outcome-context: assessing whether the Housing Subsidy Programme created assets through the ownership programme.

We had two difficulties with the review questions. Firstly, the focus on asset generation for both households and municipalities required us to combine two methods. Whereas to assess housing assets we could refer to the existing literature, to assess municipal assets we had to do new empirical work because little had been done. While these two types of assets are obviously linked, they are distinctly different issues for which a range of different assumptions exists. Secondly, each of the 14 secondary questions added a different emphasis. Although most of these questions were related, during the review process it proved difficult to devote sufficient attention to all of them. For example, the question about whether title deeds do indeed provide poor people with a platform for market access was a specific focus that required much attention – one that proved to be difficult to answer given that title deeds had to date been issued to only 50% of those households which had received a housing asset as part of the Housing Subsidy Programme. The wide range of questions necessitated a wide range of literature searches on the assumption that a considerable body of research is already available on each of the issues.

The review process

The main research question of the review was whether the Housing Subsidy Programme had provided assets to the poor and whether these assets had helped poor households to escape from poverty. The review process evolved in four phases over an originally envisaged period of 6 months. In the end, the process took more than 1 year to complete. In **Phase 1,** the DHS framed the questions in collaboration with the DPME and an evaluation steering committee, and subsequently appointed an external review team based at the University of the Free State to conduct

the review. The review team had to suggest a review method. Originally, the review team proposed the idea of conducting a systematic review to the commissioning departments. In the inception phase of the project, the limitations of the proposed methods were pointed out by the commissioning departments; the weaknesses of this approach soon became apparent in the initial literature scan conducted by the review team. Most of the literature in housing was to be found in grey literature sources and not in academic studies. The existing research also varied in design, so that while many case studies had rich qualitative data, they suffered from a lack of randomised control trials or other impact-evaluation measures – a situation often encountered in health-related research. This situation provided further justification for the review team to change the initial method and a critical realist review was thus proposed to the commissioning departments. The commissioning departments, in approving this methodology, noted that it provided the necessary flexibility and also presented a methodologically defensible approach to respond to the review questions. Phase 1 also saw the introduction of a review team to the evaluation steering committee, one that was established by the commissioning departments in line with the requirements of the National Evaluation Plan. The evaluation steering committee comprised staff from the DPME, the DHS, National Treasury, a number of officials from local municipalities and a number of hand-picked academic researchers. The DPME also appointed two external peer reviewers to comment on the work of the review team at different stages of the review process.

Phase 2: Conceptualisation and search strategy

Once the review team was familiar with the terms of reference, the team familiarised itself not only with the housing theory of change pertaining to asset building but also with the various theories of asset building. The review team had to indicate from which paradigm it would view asset building. The team argued that it largely accepted the framework of asset building portrayed by the theory of change. Yet, it was also made clear that it would adopt a more critical and normative stance in this regard. The main point is that, as reviewers, we had to work with the theory of change prescribed by the Housing Subsidy Programme. In line with the realist position that programmes should be regarded as theory, the review team at this stage also spent time with the commissioning departments in reviewing and attempting to understand the theory of change that had been developed by the commissioning departments. After this, the review team developed a detailed methodology chapter in which it set out the literature search strategy, where the search would be conducted and how the information would be synthesised. This was an expansion of what the review team had presented to the commissioning departments during the project inception phase. In line with the realist approach, the search strategy comprised a set of search terms, databases to be searched and other information. The strategy, however, allowed for the review team to use other manual search processes like reference lists of studies reviewed and word-of-mouth suggestions by experts in the field of housing, which enabled the process to remain open and flexible as new literature was found and added to the review.

Phase 3: Search process

Phase 3 was a structured literature search using not only various databases but also documents provided by the DHS. In line with the realist review method, we formulated the following 28 CMO configurations, i.e. hypotheses, directly related to the theory of change provided by the DPME and the DHS:

- Housing subsidies improve social networks and create social capital.
- Housing subsidies improve health outcomes.
- Housing subsidies improve educational outcomes.
- Housing subsidies create security of tenure for women.
- Housing subsidies create security of tenure for the aged.
- Housing subsidies create security of tenure for the disabled.
- Housing subsidies reduce mobility.
- Housing subsidies improve household stability.
- Housing subsidies result in a higher degree of citizenship responsibility.
- The Capital Housing Subsidy results in a feeling of improved security of tenure.
- Housing subsidies engender feelings of belonging.
- Housing subsidies improve social inclusiveness and integration.
- Housing subsidies result in positive attitudes towards one's own 'asset' (house).
- Housing subsidies help restore people's dignity.
- Housing subsidies allow households to trade their units.
- Housing subsidies enable households to 'climb the housing ladder'.
- Housing subsidies allow people to raise collateral for other business activities.
- Housing subsidies make it possible to obtain mortgage finance.
- Housing subsidies reduce expenditure on transport if the houses are well located.
- Housing subsidies have a positive impact on home-based enterprises.
- Housing subsidies help increase household income.
- Housing subsidies can result in rental income.
- Housing subsidies lay the foundation for increased investment in housing.
- Housing subsidies lead to an increase in the property values of units.
- The informal trading of subsidised housing units mitigates their potential value.
- Housing subsidies improve households' access to employment.
- Housing subsidies alleviate poverty.
- Housing subsidies increase poverty.

The search process we followed was iterative and flexible, and continued throughout all the stages of the review. Unlike conventional review methods in which literature searches cover a specific period and follow a strict process that is articulated in a search strategy, in our review process, literature was included as and when it came to the notice of the review team. This iteration process enriched the review process and ensured that no important seminal studies were left out of the review process.

The realist approach requires contending with four main ideological viewpoints: the neoliberal, the Marxist, the American welfare-policy view and the developing country asset-accumulation view. The neoliberal view sees housing and asset building largely in terms of the market, whereas the Marxist view is that housing should in no way be commodified. Between these two extremes, we find two main schools of thought – one originating from research on asset building in the USA, emphasising the importance of investing in housing to pay for education and retirement[25], and the other from research in developing countries, emphasising the importance of asset building for poor people in urban areas, for health, employment and stability, and particularly for stability for migrants[20].

These ideological presuppositions dominate much of the research on housing. In contrast to the practice in the health professions, housing research findings do not originate from randomised control trials but mainly from case studies, and are influenced by the researchers' ideological presuppositions. Given South Africa's apartheid past, a large portion of housing research is situated within critical theory that is known to be sceptical of markets. During a feedback workshop one person remarked that 'these academic papers originate from non-market ideologies' – a strong statement, but it does indicate the extent to which

ideology is involved in deciding whether housing has succeeded in creating assets for the poor.

In reviewing the literature we thus also had to understand the researchers' ideologies. We often had to make decisions about the value of a contribution solely on the basis of its authors' ideological presuppositions or had to take into account ideologically opposite findings. Overall, we could divide the studies into two categories: theoretically thorough work based on rather scant empirical results, and work based on large empirical data sets but theoretically shallow and moreover riddled with methodological concerns. The ideological problem was further complicated by the fact that the theory of change was based on the assumption of an ideal condition: increased access to the housing market for the poor. Table 1 shows – by means of an overview of the main findings from our sources – how we tested some of the CMO configurations.

We found approximately 1160 relevant sources with which to test our hypotheses; some sources were relevant to more than one hypothesis. The DHS also provided existing research and evaluations that they had previously commissioned. Then, we examined and assessed the titles and the available abstracts for relevance to the review questions. We found 320 research reports and papers to be relevant to the review questions. These sources included both academic and grey literature identified by means of the process described above. Towards the end of the search, we added new papers that we had found during the research process – a practice commonly followed in realist reviews. The existing research was found to have a number of shortcomings. Firstly, most of the already existing research focused on the early stages of housing development processes. Earlier studies tended to focus on variables or on the immediate outcomes of the housing development processes on households and neighbouring communities. Longer-term assessments were few. Because asset-generation is a long-term activity, the absence of long-term assessments was a major shortcoming. Evaluations over more than one generation are more likely to reflect on issues pertaining to asset building. Secondly, two paradigms of research dominate South African housing research. The first pole, critical theory, has been

Table 1: Overview of the main findings within the review framework

Key theme used in the analysis	Context	Mechanism	Outcome	Embeddedness
Access to mortgage finance and collateral	Substantial economic growth between 2001 and 2007 Global financial crisis in 2008 Negative effect of HIV/Aids on mortgage finance	Promoting access Agreements between government and banks (e.g. Record of Understanding and Financial Sector Charter); securitisation; age of settlements; locational factors Inhibiting access 40–50% of subsidised houses without title; affordability and targeted nature of the subsidy; high levels of debt	Substantial increase in number of mortgages (value ZAR500 000–1 800 000) since mid-1990s – 1.6 m household mortgages by early 2000s Large number of small mortgages provided in late 1990s (Record of Understanding) and around 2006 (Financial Sector Charter) Less than 10% of subsidised houses linked to mortgage finance and percentage in decline (becoming more difficult to provide mortgages to lower-income households) Mortgage access for old stock higher than for subsidised houses New mortgage flows stagnated since 2008 11 – 16% of subsidised houses linked to microcredit	Housing delivery process completed without provision of title to at least 40% of beneficiaries Fear of losing home Distrust of banks
Property values	Historical disenfranchisement of black people as a result of apartheid	Enhancing property values Older stock; good location Inhibiting escalation of property values Lack of market activity (formal transaction of only 11% of former township houses and 1% of subsidised houses annually); difficulties concluding transactions; lack of title; poor locations; unaffordability; sales restrictions; new houses in limited supply; declining number of houses linked to mortgages	Value of subsidised housing stock in former black townships increased since 2001 Subsidised houses generally not trading on secondary market at price comparable to cost of providing the houses Substantial evidence of housing improvement that should boost housing values Old stock and self-built houses obtain higher prices	Low-income households prefer to avert risk Infrastructure not seen by households as part of cost of housing
Climbing the housing ladder	Limited access to mortgages; housing market not functioning well Title not available to large number of beneficiaries Unaffordability	Inhibiting factors Other stock not available – household has nowhere to go if house is sold	Housing subsidies important in providing houses on first rung of housing ladder Formal transactions entered into by less than 7% of beneficiaries of subsidised houses – significantly smaller percentage than average for township houses Willingness to sell generally very low	Inherent owner scepticism about possibility of climbing housing ladder

instrumental in challenging apartheid housing policies. The second pole contains research largely based on a positivist research paradigm or in some cases 'ideologically neutral'. Housing research is generally either conceptually or theoretically rich but empirically underdeveloped or empirically rich but conceptually poor. Thirdly, the notions of *housing* and *asset accumulation* are not a prominent research direction in South Africa. Asset-based welfare or asset-based development has received scant attention in South Africa. The majority of the research on asset generation has to date originated from NGOs and individuals not affiliated to universities. The majority of the research has moreover hitherto been narrowly focused on housing as an economic asset. Asset building is also not viewed in a more holistic framework – which happens to be the conceptual framework used in the present review. Lastly, because housing research in metropolitan areas dominates the housing research landscape, we also know very little about housing issues in smaller urban settlements.

Phase 4: Hypothesis testing

In Phase 4 we used these sources to test our 12 hypotheses. We identified the specific research contexts, mechanisms and outcomes related to each source linked to a specific hypothesis. In our review, we noted the extent to which housing practice as revealed by these sources was based on specific case studies and was therefore not necessarily generalisable. Finally, the DPME asked the project team to test whether the data collected supported the existing hypotheses. Having done this, we then sought further clarification through interviews with the authors who wrote the initial texts. Because the review was part of the National Evaluation Plan that adopted utilisation-focused evaluations, the participation of users of evidence was important. The review was therefore carried out with active participation of the implementing departments and their key stakeholders, including National Treasury. Like the previous deliverables, the results from Phase 4 were presented in an evaluation report that was submitted to the DPME and the DHS for review. To test our analysis, this report was also presented in a number of workshops attended by government officials, prominent academics working in this specific field and by people in the NGO sector. Although we, as the reviewers, had a certain level of independence, the stakeholders shaped the review questions and the different outputs of the review, such as including interpretation, analysis and recommendations.

Analysis

Having provided an overview of the process, we turn to an analysis of the review method.

Working with a contested theory of change

The national housing theory of change has a number of outcomes for which there is not always consensus. Although the literature on housing and the theory of change have a number of pathways through which households that receive fully subsidised houses are able to escape poverty and build wealth, one pathway has been dominant in research and evaluation. This pathway is that which argues that a functional property market will be created through the following ways: the subsidised housing appreciates in value; subsidised houses are incorporated into the property market; subsidised houses enter municipalities' rates rolls; the value of the poor's share of the property market grows; and the poor move up the housing ladder. The dominance of this particular pathway could, theoretically, and from a measurement point of view, be ascribed to the fact that it is relatively well established in the literature. However, the theory also acknowledges several factors that block this pathway: racially skewed participation in the property market (because apartheid determined suburbs along racial lines), biased distribution of resources and wealth, high levels of poverty and unemployment, minimal private sector investment in low-income areas, and a dearth of research on how black people – with little experience of dealing in the property market because apartheid prevented black ownership of property – function in the property market. On the positive side, the theory acknowledges factors that clear the pathway: well-located land, effective planning and deeds registration, the creation of functional neighbourhoods, access to private sector finance, and good quality housing.

The anomaly between the intended outcome and the contextual limitations entailed the risk that the reviewers could easily align with either a pro-market or an anti-market perspective. Probably more problematic is the fact that some of these inhibiting factors could prove to be so overwhelming that the theory of change might not be practically possible. It also provides only a single mechanism by means of which asset building can take place, namely the housing market. However, existing research suggests a range of alternative ways of creating assets[26], such as education, settlement stability and intergenerational transfers. Focusing a theory of change only on the market does not engender a holistic understanding of assets. The review team early on pointed this fact out and the DPME and the DHS accepted a broader understanding of housing assets. This revision highlights the importance of reaching agreement on the theory of change on which the review process is focused.

The importance of review questions

Pawson and Tilley[1] argue that reviews need clear policy questions which are suited to the approach. The review questions with which we were working were not developed with the realist evaluation approach in mind and there were also too many review questions. Because the review was commissioned by government departments, there was furthermore no flexibility to adjust the review questions. Having too many questions meant that a number of CMO configurations needed to be tested. This turned out to be a challenge and we were not always able to subject the CMO configurations to thorough testing. Also, the combination of a review and questions requiring primary research made the project difficult to manage. The lesson is that even when reviews have to respond to pressing policy questions, it is important that the questions be streamlined and that the commissioners of the research should not expect the reviewers to respond to all the pressing policy questions at the same time.

Synthesising and reporting issues

Although the realist review method is theoretically sound, in practice, the analysis of the relationships between context, mechanism and outcome requires much effort. This is a limitation that Pawson and Tilley acknowledge. The idea of programmes as open systems is, for example, theoretically useful because they allow the evaluators to see the programme as part of a broader social and economic system. This idea does, however, make the boundaries of the programme wide and thus not very definitive. In our case, this meant that a wide range of articles could be considered in the review. Also, because a realist review can potentially include a range of studies with different paradigms, methods, etc., that test a number of hypotheses, it can be intellectually enormously challenging. There are no simple tick-box solutions for how findings are presented. Attempting to synthesise across more than 400 studies, testing 12 hypotheses underpinned by four theoretical/philosophical views was not always easy. Although this, in itself, is not an issue, the review team had to work through a considerable volume of data. However, this volume of data combined with answering more than one review question and also synthesising across many studies, considerably complicated our task.

Reconciling methodological values and commissioners' expectations

The commissioners of the review hoped the review would offer judgement on the effect that the Housing Subsidy Programme has had on the asset base of the poor and also on the effect that housing has had on poverty. However, this was a difficult task. The ability of the Housing Subsidy Programme to produce assets that the poor can use to help them escape poverty is contingent on so many factors that it was difficult to declare with certainty what effect the Programme had had in which context and for which category of beneficiaries. Perhaps, too, the expectation was too high. As Pawson et al.[3] argue, realist reviews can deliver only a better understanding of a programme and not general truths. For programmes as extensive and important as the South African Housing Subsidy Programme, this kind of knowledge is both appropriate and useful. The lesson is that evaluators using the realist approach need to work

at securing an understanding from the commissioners of government evaluations as to the kinds of answers the approach will generate. The commissioners moreover have to invest in unpacking and interrogating the evaluation findings to understand the implications for policy. In this case, the evaluator really walks alongside policymakers[27] and plays the role of what Pawson and Tilley refer to as 'alerting the policy community to the caveats and considerations that should inform decisions'[28].

Contested programme outcomes

Because different theoretical paradigms, analytical lenses and various academic schools of thought are involved in research on the Housing Subsidy Programme, the existing literature yields no clear-cut normative position. Although the realist approach offers some ways of dealing with this ambiguity, it nevertheless remained difficult to synthesise the body of evidence on any of the CMO configurations that were tested. Obviously, the normative positions in many of the research papers were determined by the ideological positions of the researchers. Although we attempted to factor this preference into the analysis, it did not assist in creating a better understanding. In fact, the ideological divide just became bigger. Obviously, the contested nature of housing research reflects the contested nature of housing itself. Yet, from an evaluation point of view, it remains difficult to reconcile different conclusions from the same reality. There is evidence of both market failure and of some housing asset generation. If the market has failed 80% of the people who traded, could it be a valid conclusion that market failure should necessarily lead to the abandonment of the programme?

Dealing with limitations in primary studies

Housing research in South Africa is heterogeneous and the research landscape is dominated by regional, qualitative case studies. Systematic evaluations, particularly of government programmes, are even more limited. In housing, very few studies can be classified as 'programme evaluations' and even fewer have established the effects of programmes with sufficient methodological rigour. Any form of review that discounts qualitative findings is sure to bypass the bulk of the research in the South African housing sector. We were able to use the realist review method's CMO configurations to identify some regularities and patterns across the different local case studies. However, case studies and other cross-sectional studies were not adequate to respond to the issue of asset creation as thoroughly as the contractors would have liked. The criticism that the quality of research matters in critical reviews remains important and the uneven nature of existing research in our review was problematic. This was not a weakness of the methodology itself but did point to a lack of investment in theory development in most research and to weaknesses in how housing research agendas are crafted. There is limited synergy between policy issues and the kind of research being done by academics and other partners.

Despite the challenges reflected above, the approach was useful for clarifying which factors will help the Programme to achieve the intended outcomes and also for pointing out what government should focus on to build assets for the urban poor. It was also useful for clarifying on what not to focus, for example, the focus on the poor prevents market access for poor households in a secondary market.

Conclusion and final reflections

We have shared our experience of using the realist review method to evaluate a government programme. We have explained some of the difficulties in responding to broad policy questions and the way in which this method helped us to assess a complex social programme and deal with research of an inconsistent quality (mostly small studies using qualitative methods). A realist review can help in explaining what change is happening, for whom and how, and in showing which aspects of a programme create enabling conditions for results. These attributes make it a useful framework for reviewing politically important programmes like the South African government's Housing Subsidy Programme. The government does not intend to abandon the Programme; it is a central element of the country's democracy. The review, in which we took part, was intended to strengthen elements of the Programme that are not functioning properly to enhance performance and help achieve results. The findings of the realist review alert government to those components of the Programme that need strengthening and help it to respond to a context that is complex and evolving.

This methodology has much potential in reviews and evaluations of other large, complex government interventions. However, we have pointed out that, in our case, the review was complicated by a number of issues. Firstly, the theory of change with which we were provided emphasised one pathway of change, namely a market-orientated approach to asset building while a substantial portion of asset building takes place outside market processes. Also, because this particular theory of change is contested, the findings from different studies and comments/inputs from different sector experts were sometimes irreconcilable. Secondly, too many review questions inhibited a focused review and the analysis of the relationships between context, mechanism and outcome was difficult. This was further complicated by the fact that there are many different ways in which the poor use housing to escape poverty. The latter circumstance necessitated the need to test a number of CMO configurations, which considerably complicated the synthesising and presentation of findings. Thirdly, the ideological difference between housing provision and housing research was a dominant factor in assessing the literature. In the end, it turned out that very few studies used asset generation as an important point of departure, which in its turn made reference to the review questions more difficult. Fourthly, a further challenge was that of reconciling the explanatory nature of the findings from the realist review with the commissioners' expectations of conclusive findings as regards the impact of housing on asset creation. Lastly, South African housing research is not always empirically sound and most studies tend not to address issues that are relevant to policy. As synthesis relies on existing research studies, this shortcoming created its own challenges.

However, from our experience, realism offers potential as an alternative to conventional review methods, not only in evaluation synthesis but also in primary programme evaluations. The challenges faced in this review however should not deter those who want to explore the use of realism in assessing housing programmes or any other complex programmes. We have highlighted areas in which evaluators will need to re-think to improve the application of the methodology in programme evaluations.

Acknowledgements

The Departments of Performance Monitoring and Evaluation and Human Settlements are acknowledged for funding the original review.

Authors' contributions

M.M.A. and L.M. wrote the manuscript. J.S.C. made comments on the manuscript. All authors were involved in the original review – M.M.A. as part of the contracting and L.M. and J.S.C. as service providers.

References

1. Pawson R, Tilley N. Realistic evaluation. London: Sage; 1997.
2. Grant M, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. Health Info Libr. 2009;26(2):91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x
3. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review: A new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10(1):21–34. https://doi.org/10.1258/1355819054308530
4. Ankem K. Evaluation of method in systematic reviews and meta-analyses published in LIS. Libr Inf Res. 2008;32:91–104.
5. Pawson R. Simple principles for the evaluation of complex programmes. Cidades- Comunidades e Territórios. 2004;8:95–107.
6. Marais L, Matebesi Z. Evidence-based policy development in South Africa: The case of provincial growth and development strategies. Urban Forum. 2003;24(3):357–371. https://doi.org/10.1007/s12132-012-9179-4
7. South African Department of Housing. Breaking new ground. A comprehensive plan for the development of sustainable human settlements. Pretoria: Department of Housing; 2004.

8. South African Department of Human Settlements. Housing delivery data provided by the Department of Human Settlements. Pretoria: Department of Human Settlements; 2017.

9. Westhorp G. Realist impact evaluations: An introduction [document on the Internet]. c2015 [cited 2017 Jul 17]. Available from: www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9138.pdf

10. Wong G, Greenhalgh T, Pawson R. What is a realist review and what can it do for me? An introduction to realist synthesis [presentation on the Internet]. c2009 [cited 2017 Jul 17]. Available from: http://pram.mcgill.ca/i/Wong_G_JUNE09_what_is_a_realist_review_presentation.pdf

11. Marchal B, Van Belle S, Westhorp G. Realist evaluation [webpage on the Internet]. c2015 [cited 2017 Jul 15]. Available from: http://betterevaluation.org/en/approach/realist_evaluation

12. Maxwell JA. What is realism, and why should qualitative researchers care? In: Maxwell JA, editor. A realist approach for qualitative research. Thousand Oaks, CA: Sage; 2012. p. 3–13.

13. Linsley P, Howard D, Owen S. The construction of context-mechanisms-outcomes in realistic evaluation. Nurse Res. 2015;22(3):28–34. https://doi.org/10.7748/nr.22.3.28.e1306

14. Pritchett L, Sandefur J. Context matters for size: Why external validity claims and development practice do not mix. J Global Develop. 2013;4(2):161–197. https://doi.org/10.1515/jgd-2014-0004

15. Tomlinson M. From 'quantity' to 'quality': Restructuring South Africa's housing policy ten years after. Int Develop Plan Rev. 2006;28(1):85–103. https://doi.org/10.3828/idpr.28.1.4

16. Edgely A, Stickly T, Timmons S, Meal A. Critical realist review: Exploring the real, beyond the empirical. J Further Higher Educ. 2016;40(3):316–330. https://doi.org/10.1080/03

17. Farrington D. Methodological quality standards for evaluation research. Ann Am Acad Polit Soc Sci. 2003;587(1):49–68. http://journals.sagepub.com/doi/abs/10.1177/0002716202250789

18. Amisi M, Vawda A. Strengthening democratic governance in human settlements through evaluations. In: Podems D, editor. Democracy and evaluation: Exploring the reality. Charlotte, NC: Information Age Publishing; 2017. p. 122–138.

19. Tissington K. A resource guide to housing in South Africa 1994–2010: Legislation, policy, programmes and practice. Johannesburg: Socio-Economic Rights Institute of South Africa; 2011. Available from: www.seri-sa.org/images/stories/SERI_Housing _Resource_Guide_Feb11.pdf

20. Dawson H, McLaren D. Monitoring the right of access to adequate housing in South Africa: An analysis of the policy effort, resource allocation and expenditure and enjoyment of the right to housing. Studies in Poverty and Inequality Institute working paper 8 [document on the Internet]. c2014 [cited 2016 Nov 17]. Available from: http://spii.org.za/wp-content/uploads/2014/09/Working-Paper-8_Monitoring-the-right-to-adequate-housing-in-SA.pdf

21. Huchzermeyer M. Housing for the poor? Negotiated housing policy in South Africa. Habitat Int. 2001;25:303–331. https://doi.org/10.1016/S0197-3975(00)00037-0. https://doi.org/10.1016/S0197-3975(00)00037-0

22. Charlton S, Kihato C. Reaching the poor? An analysis of the influences on the evaluation of South Africa's housing programme. In: Pillay U, Tomlinson R, Du Toit J, editors. Democracy and development: Urban policy in South Africa. Cape Town: HSRC Press; 2005. p. 252–282.

23. South African Department of Human Settlements and Department of Performance Monitoring and Evaluation. Synthesis evaluation of whether the provision of state subsidised housing addressed asset poverty for households and local municipalities? Pretoria: Department of Human Settlements and Department of Performance Monitoring and Evaluation; 2015.

24. Shisaka Development Management Services. Housing subsidy assets: Exploring the performance of government subsidised housing in South Africa. Johannesburg: FinMark Trust; 2011.

25. Sherraden M, Gilbert N. Assets and the poor: New American welfare policy. New York: M.E. Sharpe; 1991.

26. Moser C, Felton A. Intergenerational asset accumulation and poverty reduction in Guayaquil, Ecuador 1974–2004. In: Moser C, editor. Reducing global poverty: The case for asset accumulation. Washington DC: Brookings Institution Press; 2007. p. 18–34.

27. Rallis S, Rossman G. Dialogue for learning: Evaluator as critical friend. New Dir Eval. 2000;86:81–92. https://doi.org/10.1002/ev.1174

28. Pawson R, Tilley N. Realist evaluation [document on the Internet]. c2004 [cited 2018 Feb. 16]. Available from: http://www.communitymatters.com.au/RE_chapter.pdf

Productive knowledge, poverty and the entrepreneurial challenges of South African towns

AUTHOR:
Daan Toerien[1]

AFFILIATION:
[1]Centre for Environmental Management, University of the Free State, Bloemfontein, South Africa

CORRESPONDENCE TO:
Daan Toerien

EMAIL:
dtoerien@gonet.co.za

DATES:
Received: 12 Sep. 2017
Revised: 07 Dec. 2017
Accepted: 13 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
enterprise richness; entrepreneurial space; poverty trap; economic development; poor communities

HOW TO CITE:
Toerien D. Productive knowledge, poverty and the entrepreneurial challenges of South African towns. S Afr J Sci. 2018;114(11/12), Art. #4765, 8 pages. https://doi.org/10.17159/sajs.2018/4765

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
University of the Free State

Stagnant exports per capita and growing poverty in South Africa necessitated an examination of the links between the levels of productive knowledge (measured as enterprise richness), poverty (measured as Enterprise Dependency Indices) and entrepreneurial development (measured as the number of enterprises) in 188 South African towns. Two statistically significant relationships were used to examine groups of towns with different poverty levels: a linear relationship of population size and enterprise numbers, and a power law relationship of population size and enterprise richness. Increased poverty levels severely impact current and future enterprise development, despite the fact that entrepreneurial space develops similarly in wealthy and poor towns. Two broad types of entrepreneurial opportunities were discerned: starting more enterprises of types that are already present in towns, and starting enterprises of types that have not been present before. The latter requires the expansion of productive knowledge. Doubling of productive knowledge (measured as enterprise richness) more than doubles the number of enterprises in towns. The economic growth of towns always requires additional enterprises of types not yet present. This requirement is more stringent in towns with fewer than 100 enterprises, but even in large towns, enterprise growth has a Pareto-like requirement of 20% of new enterprise types. There is evidence of a 'catch-22'-like poverty trap for poor towns: they lack productive knowledge, yet to overcome poverty they need to have productive knowledge. Escaping this trap will be extremely difficult and development plans and policies should heed these findings.

Significance:

- The link between productive knowledge and the wealth/poverty status of South African towns is quantified.
- There is a 'catch-22'-like poverty trap that is difficult to escape in poorer towns.
- These findings can assist in plans to combat poverty.

Introduction

By the end of the 20th century there was a need to address growing levels of urban poverty in Africa, Latin America and much of Asia.[1] South Africa's high aggregate level of income inequality increased between 1993 and 2008, and the same is true of inequality within each of South Africa's four major racial groups. Income poverty had fallen slightly in the aggregate but it persisted at acute levels for the black African and coloured racial groups.[2] Poverty in urban areas had increased.

For more than 50 years, poverty measures in South Africa have taken distributional issues and the causes and implications of deprivation into account.[3] Most South African analyses of poverty have recognised and incorporated its multidimensional nature. Quantitative absolute measurements rely on surveys of income and consumption and on international thresholds, such as one or two dollars per day, which enable cross-country comparisons. However, poverty has also been seen as being relative and the poor as lacking the resources with which to attain a socially acceptable quality of life.[3] A quantitative link between demographic and entrepreneurial characteristics has not been used to study poverty in South African towns.

A range of regularities, which have been interpreted in terms of entrepreneurship, have been recorded in the enterprise dynamics of South African towns.[4-16] These include statistically significant linear regressions between the population and total enterprise numbers in South African towns[4-10] and can be stated as:

$$Enterprises = b(population) + C \qquad \text{Equation 1}$$

The regression coefficient b is:

$$b = \frac{Enterprises}{Population} \qquad \text{Equation 2}$$

Based on an analysis of Eastern Cape Karoo towns, Toerien[13] suggested that the inverse of the regression coefficient, $(1/b)$, which relates to how many persons are associated with the average enterprise in a group of towns (linearly correlated) is a measure of the wealth/poverty status of the group of towns. It is called the Enterprise Dependency Index (EDI):

$$\frac{1}{b} = EDI = \frac{Population}{Enterprises} \qquad \text{Equation 3}$$

More persons per enterprise in a town indicates more poverty, and fewer persons per enterprise, wealthier conditions.

Equation 3 can be restated as:

$$Enterprises = \frac{Population}{EDI} \qquad \text{Equation 4}$$

The enterprises in towns are thus related to the magnitude of their populations as well as their wealth/poverty status (indicated by EDI).

Changes in the population of towns are the result of the net growth rate (birth rate minus death rate), in-migration and out-migration. By the time of the first post-apartheid census in 1996, just over half of the South African population lived in urban areas; this number grew to 57.5% by 2001 as there was a movement of people to cities experiencing economic growth.[17] However, in the Gauteng Province, South Africa's dominant migration destination, some 70% of the population growth between 1996 and 2001 was the consequence of natural increases.[18] Natural processes and migration have influenced the population dynamics of South African towns.

The enormous income gaps between rich and poor nations are an expression of the vast differences in productive knowledge amassed by different nations.[19] The differences are expressed in the diversity and sophistication of the products of each of these nations. The social accumulation of productive knowledge has not been a universal phenomenon. It has taken place in some parts of the world, but not in others. Where it has happened, it has underpinned an incredible increase in living standards. Where it has not, living standards resemble those of centuries past.[19] If the level of the productive knowledge of countries determines the economic fates of the countries and their populations, the same should be true for local economies and populations of towns.

A statistically significant log-log relationship (hereafter called a power law) has been recorded between the total enterprise numbers and the number of enterprise types (referred to as enterprise richness, ER) in a large group of South African towns.[15] This relationship has endured over approximately 70 years in a selection of Karoo towns.[16] The economic growth of any South African town, small or large, is therefore dependent on new business ideas and the start-up of enterprise types not yet present in (i.e. new to) the town (Figure 1). As towns grow, a Pareto-like (80:20) division is reached between enterprise types already present and new enterprise types not yet been present in the towns.[15,16]

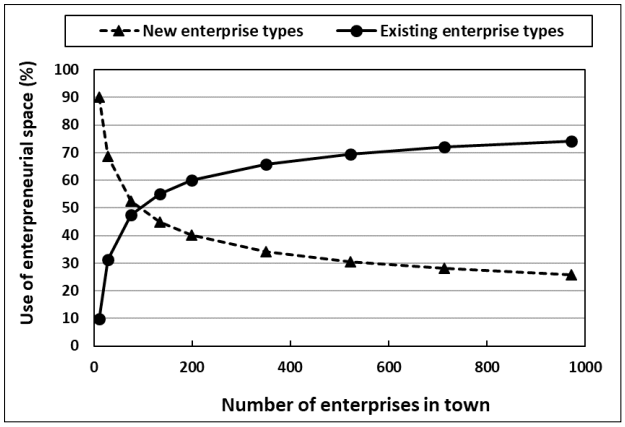


Figure 1: The use of entrepreneurial space in 188 South African towns by two groups of entrepreneurs: those founding new enterprise types and those founding more of existing enterprise types.

An increased ER in a town is, therefore, a direct indication of a higher level of productive knowledge among its residents, and hence its 'entrepreneurial capacity'. Schumpeter[20] said about 'creative destruction':

> [T]*he same process of industrial mutation – if I may use that biological term – that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one.*

Hausmann and Klinger[21] argued that producing new things is quite different from producing more of the same. Florida[22] remarked:

> *Human creativity is the ultimate economic resource. The ability to come up with new ideas and better ways of doing things is ultimately what raises productivity and thus living standards.*

The entrepreneurial well-being of South African towns is clearly connected to the ability to conceive and deliver new products and/or services. ER may serve as a proxy measure of productive knowledge.

The ER relationship can now be restated as:

$$ER = A(Enterprises)^e \qquad \text{Equation 5}$$

where A is a constant and e is a coefficient.

Incorporating Equation 4 into Equation 5 results in:

$$ER = A(Population/EDI)^e \qquad \text{Equation 6}$$

The reduction of poverty and inequality remain significant problems in South Africa. According to Equation 6, ER is related to the population size of towns and their wealth/poverty status. Based on the ideas of Hausmann et al.[19], higher levels of productive knowledge should, therefore, be quantitatively related to wealthier towns (lower EDIs) and lower levels to poorer towns (higher EDIs). Quantification of the impacts of poverty on entrepreneurial development in South African towns would provide a new way of analysing the poverty problem.

The prime purpose of this contribution is, therefore, to investigate the quantitative links between productive knowledge (measured as ER) and the wealth/poverty status of South African towns (measured as EDIs). In the process, future scenarios of the entrepreneurial evolution of towns of different wealth/poverty status are used to sketch the debilitating impacts of poverty.

Productive knowledge

Wealth and development are related to the complexity that emerges from the interactions between the increasing number of individual activities that constitute an economy.[23] Based on these ideas, Hausmann and co-workers produced the *Harvard MIT Atlas of Economic Complexity and Maps of Paths to Prosperity*.[19] It is based on data extracted from 128 countries representing 99% of world trade. A central tenet of their conclusions is that the differential accumulation of productive knowledge distinguishes between rich and poor countries. These differences are expressed in the diversity and sophistication of the things that each of these nations makes, or in other words, their abilities to produce products and services that have uniqueness. They concluded that productive knowledge to create new products or services is key to economic success and wealth.

Productive knowledge is not available in books or on the Internet but is embedded in brains and human networks. It is tacit and hard to transmit and acquire. It comes more from years of experience than from years of schooling.[19] Hausmann and Klinger[21] suggested that South Africa's stagnant exports per capita over the past 40 years is a consequence in part of the peripheral nature of its productive capabilities: the country is specialised in sectors intensive in highly specific factors of production that cannot be easily redeployed to other activities.[1] In other words, South Africa lacks in productive knowledge.

It is argued here that: (1) the entrepreneurial well-being of South African towns is clearly connected to the abilities of their residents to conceive and deliver new products and/or services, or in other words, connected to the levels of their productive knowledge, (2) ER can be used as a proxy measurement of productive knowledge, and (3) the concepts of ER and productive knowledge provide new ways to examine the socio-economic dynamics of South African towns.

Methods

Experimental design

The links between poverty, productive knowledge and entrepreneurship in South Africa, and internationally, are poorly studied. It is, therefore, necessary to describe how such an investigation was approached in this study. Almost 200 South African towns for which the necessary data were available were selected for the study. The presence of a statistically significant power law between ER and enterprise numbers (see Equation 5) for the total group was confirmed to ensure that the data would support this line of investigation. The resulting power law provided a reference line to examine the entrepreneurial impacts of different poverty levels. The towns were then ranked on the basis of their EDIs and divided into five groups. For the towns in each group it was confirmed that: (1) there was a statistically significant linear relationship between population numbers and enterprise numbers (see Equation 1), (2) the distribution of the EDIs of the towns in each group was not skewed and an average EDI could represent the group, and (3) there was a statistically significant power law relationship between ER and population numbers (according to Equation 6). It was previously shown that the power law relationship between total enterprises and ER has endured over almost 70 years.[16] The possibility that poverty might influence the characteristics of the power law relationship between enterprises and ER (according to Equation 5) was also investigated. Finally, different population growth scenarios were used to sketch the impacts of wealth and poverty on towns.

Selection of towns

A selection of 188 towns was used in this investigation (Tables 1 and 2). The selected towns represent all of the towns with at least 10 enterprises (as recorded in a database) at the time of writing this contribution and represent a broad range of South African towns including towns from most provinces, towns of the former homelands, towns from strong agricultural areas, towns from mining areas, etc. A range of town sizes is also represented (Table 2) but villages with fewer than 10 enterprises were excluded to avoid potential distortions in the analysis.

The towns were ranked on the basis of their EDIs and based on the magnitude of their EDI divided into five groups representing different levels of wealth/poverty. Four groups were about the same in number. The fifth group, representing the poorest towns, had only 19 towns. The groups are: Group 1 – EDIs from 10 to 80 (wealthiest towns); Group 2 – EDIs between 80 and 140; Group 3 – EDIs from 140 to 200; Group 4 – EDIs between 200 and 300; Group 5 – EDIs $\geq$ 300 (poorest towns).

Each of the groups had a spread of population sizes as judged by their minimum and maximum population numbers (Table 2). This spread ruled out the possibility that differences in EDIs might be solely a function of population size. Population sizes had reasonably skewed distributions as judged by the differences between average and median values. This was not the case as far as EDIs were concerned. The average and median EDI values were fairly close, suggesting that the EDIs within groups had reasonably normal distributions. The average EDI could therefore be used to represent each group.

Table 1: Groups and towns analysed

EDI = 10–80, *n* = 45		EDI = 80–140, *n* = 45		EDI = 140–200, *n* = 39		EDI = 200–300, *n* = 41		EDI ≥300, *n* = 19
Albertinia	Montagu	Aliwal North	Merweville	Aberdeen	Hopetown	Barkly West	Memel	Allanridge
Barrydale	Mtubatuba	Augrabies	Middelburg (EC)	Ashton	Jacobsdal	Bloemhof	Noupoort	Arlington
Bethlehem	Napier	Brandvlei	Mookgophong	Badplaas	Kenhardt	Boshof	Oranjeville	Botshabelo
Bonnievale	Nieu Bethesda	Calvinia	Parys	Beaufort West	Komga	Bultfontein	Pearston	Dealesville
Bredasdorp	Nieuwoudtville	Carnarvon	Philippolis	Bethal	Murraysburg	Clocolan	Petrusburg	Dewetsdorp
Caledon	Orania	Colesberg	Phuthaditjhaba	Bethulie	Postmasburg	Deneysville	Phalaborwa	Edenville
Calitzdorp	Oudtshoorn	Cradock	Pofadder	Bothaville	Prieska	Edenburg	Philipstown	Excelsior
Clarens	Porterville	De Aar	Richmond	Brandfort	Reddersburg	Fort Beaufort	Rouxville	Hertzogville
De Rust	Prince Albert	Fraserburg	Robertson	Britstown	Reitz	Fouriesburg	Taung	Lindley
Dullstroom	Queenstown	Garies	Sannieshof	Burgersdorp	Schweizer-Reneke	Griekwastad	Theunissen	Odendaalsrus
Gansbaai	Riversdale	Graaff Reinet	Sasolburg	Christiana	Senekal	Hanover	Tweespruit	Paul Roux
Gariepdam	Riviersonderend	Groblershoop	Somerset East	Daniëlskuil	Smithfield	Hennenman	Ventersburg	Petrus Steyn
Great Brak River	Still Bay	Heidelberg	Steytlerville	Douglas	Springfontein	Jagersfontein	Venterstad	Petrusville
Greyton	Struis Bay	Hotazel	Thabazimbi	Fauresmith	Steynsburg	Jan Kempdorp	Viljoenskroon	Rosendal
Harrismith	Sutherland	Kathu	Thohoyandou	Ficksburg	Strydenburg	Kestell	Villiers	Steynsrus
Hartswater	Swellendam	Keimoes	Trompsburg	Frankfort	Stutterheim	Klipplaat	Virginia	Thaba 'Nchu
Jansenville	Tulbagh	Kroonstad	Upington	Heilbron	Tarkastad	Koffiefontein	Warden	Tweeling
Kakamas	Uniondale	Lady Frere	Victoria West	Hendrina	Wakkerstroom	Koppies	Warrenton	Villiers
Kamieskroon	Vanderkloof	Ladybrand	Vrede	Hofmeyr	Zastron	Luckhoff	Wepener	Vredefort
Kleinmond	Vosburg	Laingsburg	Welkom	Hoopstad		Marquard	Wesselsbron	
Ladismith	Vredendal	Lime Acres	Williston				Winburg	
Loxton	Yzerfontein	Loeriesfontein	Willowmore					
Lutzville		McGregor						

EDI, Enterprise Dependency Index (population needed to 'carry' the average enterprise).

Table 2: The population and Enterprise Dependency Index (EDI) characteristics of the selected groups of towns

		Population				EDI			
EDI Group	**Number**	**Minimum**	**Maximum**	**Average**	**Median**	**Minimum**	**Maximum**	**Average**	**Median**
Total	188	892	211 011	18 473	9099	13.8	1074.3	169.6	148.3
10–80	45	892	76 667	11 675	6372	13.8	79.4	56.8	57.2
80–150	45	1592	211 011	25 017	9680	80.2	139.5	106.6	105.3
150–200	38	2987	71 011	17 099	13 112	143.0	195.1	170.3	170.5
200–300	41	2967	109 468	17 273	11 260	200.1	289.6	235.4	234.7
≥300	19	3935	181 712	24 407	9423	301.9	1074.3	442.6	357.7

Enterprises, enterprise richness and Enterprise Dependency Indices

The enterprises and enterprise types of each of the 188 towns were identified, classified and enumerated as previously described.[4,15,16] Telephone directories (supplemented by Internet searches where necessary) were used for the identification of the enterprises in each town[4] and the enterprises were then enumerated. Enterprise types were determined from a database of more than 500 enterprise types hitherto encountered in South African towns to provide the ER of each town.[15,16]

The towns were allocated to their respective EDI groups and two relationships were determined for each group: (1) the linear relationship between population size and enterprise numbers (see Equation 1), and (2) the power law between ER and the population numbers of the towns (see Equation 6). Microsoft Excel software was used for the calculations.

Wealth/poverty and entrepreneurial spaces

The power laws between ER values and the population numbers (see Equation 6) of the towns of each EDI group were calculated and compared to assess whether wealth/poverty impacts the development of entrepreneurial spaces in towns. The spread of the data points of groups was graphically examined for this purpose.

Scenarios of towns' growth

Scenarios were selected to investigate how towns of different wealth/poverty statuses would respond entrepreneurially to population growth over time. The linking of EDIs (wealth/poverty) and ER (proxy for productive knowledge) was investigated using the linear regression equations of the selected groups of towns to predict enterprise numbers for hypothetical population numbers and using power laws to predict the number of enterprise types for the same hypothetical population numbers in the scenarios. Two growth scenarios (2% per annum and 4% per annum) and two initial population sizes (10 000 and 50 000) were used. To ensure that the growth scenarios were reasonable, a frequency distribution was calculated of the population growth rates between the censuses of 2001 and 2011 of South African towns which had fewer than 200 000 residents in 2001.[24] The median population size of the towns used in this study was 9030. The selection of initial population sizes of 10 000 and 50 000 for the scenarios was considered to be representative of the 'average South African town' in the former case and 'reasonably large towns' in the latter.

Results

Population numbers and enterprise numbers of groups

The linear relationships between the population numbers (independent variables) and enterprise numbers (dependent variables) in the different groups are presented in Table 3. The relationships were all statistically significant at $p<0.01$. There was as expected a gradual lowering of the regression coefficients as the EDIs of groups increased. The group EDIs (=inverses of these coefficients, i.e. persons per enterprise) clearly separated the towns into richer and poorer representatives, thereby enabling comparisons of the impacts of wealth/poverty.

Population numbers and enterprise richness of groups

Previous studies revealed statistically significant relationships between the total enterprise numbers and the number of enterprise types (enterprise richness) in South African towns.[15,16] This study revealed for the first time that there are also statistically significant power law relationships ($p<0.01$) between the population numbers and the enterprise types in towns for each of the groups investigated (Table 3). In contrast to the progressive reduction in linear regression coefficients with increasing poverty levels, the exponential coefficients of the power laws did not show systematic change with increasing poverty levels of the groups (Table 3).

Linking of wealth/poverty and enterprise richness

The frequency distribution of growth rates between 2001 and 2011 of South African towns with more than 15 000 residents is presented in Figure 2. The growth rates are reasonably normally distributed with a peak between 2% and 3% per annum. For the scenarios used here, two growth rates (2% p.a. and 4% p.a.) were selected as being representative of typical (2%) and higher (4%) population growth rates for South African towns.

Table 3: The linear population:enterprises relationship and the power law population:enterprise types relationship of groups of towns with different Enterprise Dependency Indices (EDIs)

		Linear: Population–enterprises			Power law: Population–enterprise richness		
EDI Group	*n*	**Correlation**	**Regression coefficient**	**Intercept**	**Correlation**	**Coefficient**	**Constant**
10–80	45	0.97	0.0136	38.1	0.93	1.2671	42.9
80–140	45	0.98	0.0084	14.3	0.97	1.4422	41.2
140–200	38	0.99	0.0065	-7.3	0.98	1.2378	123.5
200–300	41	1.00	0.0049	-7.7	0.97	1.1944	195.9
>300	19	0.89	0.0012	17.6	0.91	1.4539	143.3

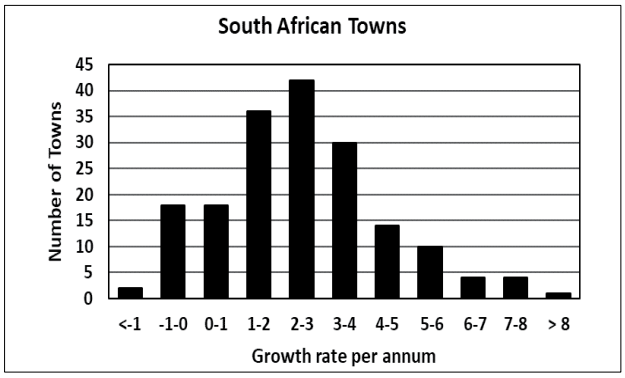


Figure 2: Frequency distribution of population growth rates of South African towns between 2001 and 2011.

It is clear that the poverty status of towns as indicated by rising EDIs has a severe influence on enterprise development, both in terms of total enterprises as well as on the numbers of enterprise types (Table 4). For instance, a rich town (EDI<80) with an initial population of 50 000 residents will have 718 total enterprises and 210 enterprise types. A similarly sized poor town (EDI>300) will have only 79 total enterprises and 47 enterprise types. After 5 years at a growth rate of 2% per annum, the former town will have 774 total enterprises and 222 enterprise types and the latter town only 85 total enterprises and 49 enterprise types. The entrepreneurial challenges of poor towns are clearly different from those of wealthier towns.

Table 5 quantifies the entrepreneurial challenges of towns with different wealth/poverty statuses and different growth rates. As towns grow there are two broad types of opportunities for entrepreneurs to start enterprises: (1) create more of the enterprise types that are already present in these towns, and (2) introduce types that have not yet been successfully started in the town. Larger towns that grow more rapidly will have more of both types of opportunities than smaller towns. These opportunities are severely impacted by poverty as is illustrated in the 10 years column in Table 5.

The prime purpose of this contribution was to investigate if the wealth/poverty status of South African towns is related to their enterprise richness, an indicator of the level of their productive knowledge pools. The results presented in Tables 4 and 5 clearly support this contention. In all scenarios, richer towns are linked to higher numbers of enterprises and more enterprise types. Importantly, the growth of all towns, irrespective of their wealth/poverty status and population sizes, requires entrepreneurs that can 'visualise' new opportunities. In this regard, richer towns benefit more than poorer towns, larger towns benefit more than smaller towns, and towns with higher growth rates benefit more than towns with lower growth rates.

Table 4: Projected increases in total enterprise numbers and the number of enterprise types in towns of two different population sizes and two different growth rates. Year 1 represents the initial conditions. Existing enterprise types = the difference between total enterprise numbers and the number of enterprise types. Higher Enterprise Dependency Indices (EDI) indicate higher poverty levels.

EDIs	Population growth rate	Initial population size	New enterprise types			Total enterprises			Existing enterprise types		
			Year 1	Year 5	Year 10	Year 1	Year 5	Year 10	Year 1	Year 5	Year 10
10–80	2	10000	70	74	79	173	184	200	103	110	121
80–140	2	10000	45	48	51	100	107	117	55	59	66
140–200	2	10000	35	37	40	58	63	71	23	26	31
200–300	2	10000	27	29	31	41	45	51	14	16	20
≥300	2	10000	19	21	21	30	31	32	11	10	11
10–80	4	10000	70	78	89	173	196	231	103	118	142
80–140	4	10000	45	50	57	100	114	136	55	64	79
140–200	4	10000	35	40	46	58	69	86	23	29	40
200–300	4	10000	27	31	36	41	50	62	14	19	26
≥300	4	10000	19	21	23	30	32	35	11	11	12
10–80	2	50000	210	222	238	718	774	851	508	552	613
80–140	2	50000	130	137	146	435	470	517	305	333	371
140–200	2	50000	121	129	139	320	347	383	199	218	244
200–300	2	50000	97	104	112	237	257	285	140	153	173
≥300	2	50000	47	49	52	79	85	92	32	36	40
10–80	4	50000	210	234	268	718	834	1006	508	600	738
80–140	4	50000	130	144	164	435	506	612	305	362	448
140–200	4	50000	121	137	159	320	375	458	199	238	299
200–300	4	50000	97	110	129	237	279	328	140	169	199
≥300	4	50000	47	52	58	79	90	106	32	38	48

EDI = population/total enterprises

Table 5: Projected net increases in total enterprises and enterprise types in South African towns subject to different growth rates and population sizes

EDIs	Population growth rate	Initial population size	New enterprise types		Existing enterprise types	
			Net growth after		Net growth after	
			5 years	10 years	5 years	10 years
25–80	2	10 000	4	9	7	18
80–140	2	10 000	3	6	4	11
140–200	2	10 000	2	5	3	8
200–300	2	10 000	2	4	2	6
≥300	2	10 000	2	2	0	0
25–80	4	10 000	8	19	15	39
80–140	4	10 000	5	12	9	24
140–200	4	10 000	5	11	6	17
200–300	4	10 000	4	9	5	12
≥300	4	10 000	2	4	0	1
25–80	2	50 000	12	28	44	105
80–140	2	50 000	7	16	28	66
140–200	2	50 000	8	18	19	45
200–300	2	50 000	7	15	13	33
≥300	2	50 000	2	5	4	8
25–80	4	50 000	24	58	92	230
80–140	4	50 000	14	34	57	143
140–200	4	50 000	16	38	39	100
200–300	4	50000	13	32	29	59
≥300	4	50000	5	11	6	16

EDI, Enterprise Dependency Index = (population/total enterprises)

Productive knowledge and total enterprise numbers

One issue remains to be resolved. The regression coefficients of linear regression equations relating total enterprises to population sizes are clearly different from one another (Table 3). However, the power law equations relating population sizes and enterprise types did not show systematic differences (Table 3). Yet the wealth/poverty status of towns clearly influences the number of enterprise types in towns (Tables 4 and 5). Against the background of an enduring relationship between enterprise numbers and enterprise types[16], similarities in the development of entrepreneurial space in wealthy and poor towns were investigated. In Figure 3 the spread of the data points of the different groups is superimposed on the power law line calculated from all of the towns used in the study. It is clear that the spread of the data points of the different groups is very similar, i.e. the development of entrepreneurial space is not affected by the wealth/poverty status of towns.

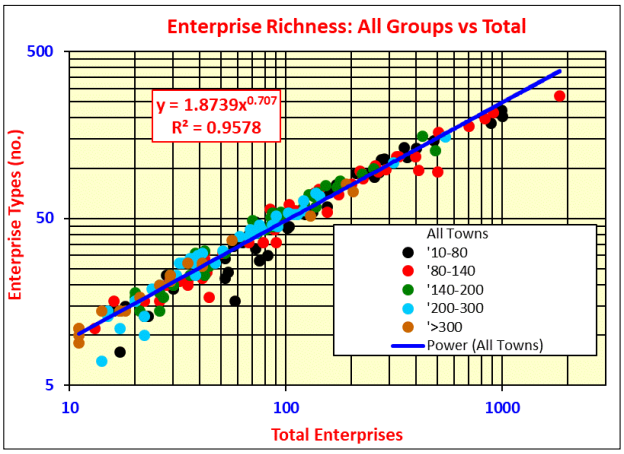


Figure 3: The relationships between total enterprises and enterprise types of the different EDI groups in relation to the power law line of total enterprises versus enterprise types of all the selected towns.

Based on the ideas of Hausmann et al.[19] that economic complexity is expressed in the composition of productive output, it was postulated (see earlier) that the ER of South African towns might be a measure of the productive knowledge embedded in them. In communities with higher levels of productive knowledge, there is a higher chance of the presence of people able to discern business opportunities that can be realised by innovative new combinations of knowledge, skills and other inputs from the community. This reasoning means that productive knowledge could be the driver of the enterprise richness/total enterprises relationship, and not vice versa. Figure 4 incorporates this possibility and presents the productive knowledge/enterprises power law relationship for the towns selected for this study (line in the graph) as well as the distribution of the data points of the different groups in relation to the line.

This power law equation:

$$\text{Enterprises in town} = 0.513(\text{number of enterprise types in town})^{1.3548}, \quad \text{Equation 7}$$

with $r=0.98$ and $n=188$, is statistically highly significant ($p<0.01$). It indicates that for each doubling of ER (i.e. the doubling of productive knowledge), the total enterprise numbers will increase by 2.56 times. This quantifies the expansion of entrepreneurial space, measured in the total number of enterprises when productive knowledge increases in towns. Importantly, this relationship is not affected by the wealth/poverty status of towns (Figure 4). Doubling the productive knowledge of poor towns will also expand the total enterprises by 2.56 times.

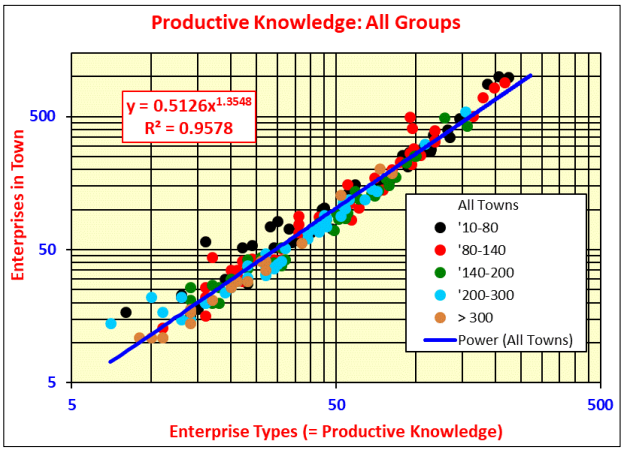


Figure 4: Productive knowledge, measured by the enterprise richness of 188 South African towns, as a driver of the enterprise types–total enterprises relationship.

Equation 7 was used to calculate the entrepreneurial challenges of South African towns in terms of new enterprise types and existing enterprise types (Figure 1). Up to a size of approximately 100 enterprises per town, the growth of towns requires more new enterprise types than existing enterprise types. The linear regression equations of Table 3 indicate that this point will be reached at approximately 7000 persons in the richer towns but only at about 80 000 persons in the poorest towns. This illustrates the impact of the presence of wealth in the population of a town and highlights the 'catch-22'-like poverty trap of poor towns: they are poor because they lack the productive knowledge to produce products or deliver services needed outside their domains. However, to overcome poverty, they need to have productive knowledge!

It must also be kept in mind that the growth of all towns, irrespective of their size, requires entrepreneurs with productive knowledge to start new types of enterprises as well as more enterprises of types already present (Figure 1). Even in very large towns, further growth is dependent on a Pareto-like division between new and existing enterprise types, i.e. about 20% of new enterprises have to be of types that have not yet been present in these towns (requiring higher levels of productive knowledge than before) and 80% must be of types already present in the town (the productive knowledge for this is already present).

Discussion and conclusions

South Africa's exports per capita over the past 40 years have been stagnant because the country is specialised in sectors intensive in highly specific factors of production that cannot be easily redeployed to other activities.[21] Hausmann and co-workers[19] state that countries tend to converge to the level of income that can be supported by the know-how that is embedded in their economies – their so-called productive knowledge. Richer countries have more productive knowledge than poor countries, and vice versa. South Africa has an obvious lack of productive knowledge to move beyond the constraints mentioned above.

Despite a general decline in poverty between 2006 and 2011, poverty levels in South Africa rose in 2015.[25] More than half of South Africans are poor. The National Development Plan[26] considers the development of entrepreneurship to be one of the key developmental issues in overcoming poverty although it admits that early-stage entrepreneurial activity rates in South Africa are about half of what they are in other developing countries.

The possibility of a relationship between poverty and the lack of productive knowledge in South African towns was investigated here. EDIs (i.e. persons per enterprise) proved to be a useful measure of the wealth/poverty status of towns and it clearly separated South African towns into richer and poorer groups (Table 3). For the first time, statistically significant power law relationships between the population numbers and enterprise types of South African towns, including richer and poorer groups, were recorded (Table 4). Similarly sized populations in richer towns can 'carry' many more enterprises than those in poorer towns. Entrepreneurial dynamics and the wealth/poverty status of South African towns are clearly linked. It will be very difficult for poor towns to escape the poverty trap. The ideas of Hausmann and colleagues about the importance of productive knowledge in the economic fate of countries also seem to apply to the economic fate of towns in South Africa.

Previous studies of enterprise richness considered total enterprise numbers as the driver of the total enterprises/enterprise richness power law.[15,16] The present contribution indicates the possibility that enterprise richness could be the driver of an enterprise richness/total enterprises relationship. This statistically significant power law (Figure 4) implies that for each doubling of the number of enterprise types (equals a doubling of productive knowledge), the total enterprise numbers will increase by 2.56 times. This quantifies how entrepreneurial space expands when enterprise richness (=productive knowledge) increases in towns.

Entrepreneurial space develops similarly in wealthy and poor towns (Figures 3 and 4). The growth of towns, irrespective of their wealth/poverty status and their population sizes, requires entrepreneurs who can 'visualise' new opportunities (Figure 1). Two broad types of entrepreneurial opportunities were discerned: (1) starting more enterprises of types that are already present in towns (existing types), and (2) starting enterprises of types not yet present in the towns (new types) (Table 5, Figure 1). The latter requires the expansion of productive knowledge. Larger towns that grow more rapidly will have more of both types of opportunities. Scenario projections indicate that richer towns are linked to higher numbers of enterprises and higher ER values, now and in the future (Tables 4 and 5). Towns with fewer than 100 enterprises will be particularly challenged to start more enterprises of new types than of existing types, and this requires higher levels of productive knowledge. As a consequence, there is a 'catch-22'-like poverty trap for poor towns: they lack productive knowledge, yet to overcome poverty, they need to have productive knowledge. However, even in large towns there is a Pareto-like requirement: 20% of new enterprises must be new types and 80% existing types (Figure 1).

The identity of new enterprise types obviously changes as towns grow larger. For example, in small towns a new enterprise type might occur when the first attorney starts a practice in a town and in larger towns it might be when the first enterprise starts offering back-office services. Examples of existing enterprises might be the second hotel in a town or the second general trader. This aspect requires further research to establish if specific patterns of enterprise development can be discerned.

Two perspectives are available to improve our understanding of the influences of poverty and the lack of productive knowledge on entrepreneurship in, and the growth of, South African towns. In the first, increasing poverty in towns induces a 'slide' down the slope of the line in Figure 3. A poorer town with the same population as a richer town will have less productive knowledge, fewer enterprises and fewer employment opportunities. In the second, towns with fewer enterprise types (i.e. less productive knowledge) will 'slide' down the slope of the line in Figure 4, resulting in fewer enterprises and fewer employment opportunities. These perspectives have potentially crucial implications for local economic development plans to reduce poverty and enhance employment.

Local economic development is a core local government mandate in South Africa.[27] However, despite the significant support it received for nearly 20 years, results have been modest. Nel and Rogerson[27] expressed concern about a potential over-focus on pro-poor local economic development. The 'catch-22'-like trap of poor communities implies that pro-poor local economic development could only be successful if the productive knowledge of these communities is increased. The insightful views of Hausmann and co-workers[19] and the results of this study should be considered when local economic development strategies to address poverty are developed and implemented.

Acknowledgements

The University of the Free State provided research funding and the Centre for Environmental Management, University of the Free State, provided administrative and research support. Linda Retief provided language editing. Alumnus services of the Massachusetts Institute of Technology provided online scholarly journal access. Marie Toerien and Estelle Zeelie provided technical assistance and two anonymous reviewers provided valuable comments.

References

1. Anzorena J, Bolnick J, Boonyabancha S, Cabannes Y, Hardoy A, Hasan A, et al. Reducing urban poverty; some lessons from experience. Environ Urban. 1998;10(1):167–186. https://doi.org/10.1177/095624789801000106
2. Leibbrandt M, Woolard I, Finn A, Argent J. Trends in South African income distribution and poverty since the fall of apartheid. OECD social, employment and migration working paper no. 101. Paris: Organisation for Economic Co-operation and Development; 2010.
3. May J. Smoke and mirrors? The science of poverty measurement and its application. Dev South Afr. 2012;29(1):63–75. https://doi.org/10.1080/03768835X.2012.645641
4. Toerien DF, Seaman MT. The enterprise ecology of towns in the Karoo, South Africa. S Afr J Sci. 2010;106(5/6):24–33. http://dx.doi.org/10.4102/sajs.v106i5/6.182

5. Toerien DF, Seaman MT. Ecology, water and enterprise development in selected rural South African towns. Water SA. 2011;37(1):47–56. https://doi.org/10.4314/wsa.v37i1.64106

6. Toerien DF, Seaman MT. Regional order in the enterprise structures of selected EC Karoo towns. S Afr Geogr J. 2012;94(2):1–15. http://dx.doi.org/10.1080/03736245.2012.742782

7. Toerien DF, Seaman MT. Proportionality in enterprise development of South African towns. S Afr J Sci. 2012;108(5/6):38–47. http://dx.doi.org/10.4102/sajs.v108i5/6.588

8. Toerien DF, Marais L. Classification of South African towns revisited. In: Donaldson R, Marais L, editors. Small town geographies in Africa: Experiences from South Africa and elsewhere. New York: Nova Science Publishers; 2012. p. 3–19.

9. Toerien DF. Enterprise proportionalities in the tourism sector of South African towns. In: Kasimoglu M, editor. Visions of global tourism industry: creating and sustaining competitive strategies. Rijeka: Intech; 2012. p. 113–138. http://dx.doi.org/10.5772/37319

10. Toerien DF. New utilization/conservation dilemmas in the Karoo, South Africa: Potential economic, demographic and entrepreneurial consequences. In: Ferguson G, editor. Arid and semi-arid environments: Biogeodiversity, impacts and environmental challenges. New York: Nova Science Publishers; 2015. p. 79–123.

11. Toerien DF, Seaman MT. Evidence of island effects in South African enterprise ecosystems. In: Mahamane A, editor. The functioning of ecosystems. Rijeka: Intech; 2012. p. 229–248. http://dx.doi.org/10.5772/36641

12. Toerien DF, Seaman MT. Paradoxes, the tyranny of structures and enterprise development in South African towns. Presented at: Strategies to overcome poverty and inequality: Towards Carnegie3; 2012 Sep 3–7; Cape Town, South Africa. Available from: http://carnegie3.org.za/docs/papers/269_Toerien_Paradoxes,%20the%20tyranny%20of%20structures%20and%20enterprise%20development%20in%20SA%20towns.pdf

13. Toerien DF. 'n Eeu van orde in sakeondernemings in dorpe van die Oos-Kaapse Karoo [A century of order in the enterprises of the towns of the Eastern Cape Karoo]. LitNet Akad. 2014;11(1):330–371. Afrikaans.

14. Toerien DF. Economic value addition, employment, and enterprise profiles of local authorities in the Free State, South Africa. Cogent Soc Sci. 2015;1:1054610, http://dx.doi.org/10.1080/23311886.2015.1054610

15. Toerien DF, Seaman MT. Enterprise richness as an important characteristic of South African towns. S Afr J Sci. 2014;110(11/12), Art. #2014-0018, 9 pages. http://dx.doi.org/10.1590/sajs.2014/20140018

16. Toerien DF. The enduring and spatial nature of the enterprise richness of South African towns. S Afr J Sci. 2017;113(3/4), Art. #2016-0190, 8 pages. https://doi.org/10.17159/sajs.2017/20160190

17. Todes A, Kok P, Wentzel M, Van Zyl J, Cross C. Contemporary South African urbanisation dynamics. Presented at: UNU-WIDER Conference: Beyond the tipping point. African development in an urban world; 2008 June 21; Cape Town, South Africa.

18. Cross C, Kok P, Wentzel M, Tlabel K, Weir-Smit G, Mafukidze J. Poverty pockets in Gauteng. How poverty impacts migration. Human Sciences Research Council Report to the Gauteng Inter-sectoral Development Unit. Pretoria: Human Sciences Research Council; 2005.

19. Hausmann R, Hidalgo CA, Bustos S, Coscia M, Chung S, Jimenez J, et al. The atlas of economic complexity: Mapping paths to prosperity. Cambridge, MA: Center for International Development, Harvard University; 2017.

20. Schumpeter JA. Capitalism, socialism and democracy. 3rd ed. New York: Harper Colophon; 1942.

21. Hausmann R, Klinger B. South Africa's export predicament. Econ Transition. 2008;16(4):609–637. https://doi.org/10.1111/j.1468-0351.2008.00337.x

22. Florida R. The rise of the creative class. New York: Basic Books; 2004.

23. Hidalgo C, Hausmann R. The building blocks of economic complexity. Proc Natl Acad Sci USA. 2009;106(26):10570–10575. https://doi.org/10.1073/pnas.0900943106

24. Citypopulation – statistics, maps & charts [webpage on the Internet]. No date [cited 2017 Sep 12]. Available from: https://www.citypopulation.de/SouthAfrica-UA.html?cityid=12488

25. Statistics South Africa (Stats SA). Poverty trends in South Africa: An examination of absolute poverty between 2006 and 2015. Pretoria: Stats SA; 2017.

26. South African National Planning Commission. Our future – make it work. National Development Plan 2030. Pretoria: National Planning Commission; 2011.

27. Nel E, Rogerson CM. The contested trajectory of applied local economic development in South Africa. Local Econ. 2016;311(2):109–123.

Modelling the length of time spent in an unemployment state in South Africa

AUTHORS:
Jeanette Z. Nonyana[1]
Peter M. Njuho[1]

AFFILIATION:
[1]Department of Statistics, University of South Africa, Johannesburg, South Africa

CORRESPONDENCE TO:
Jeanette Nonyana

EMAIL:
znonyana@randwater.co.za

DATES:
Received: 26 Jan. 2018
Revised: 06 June 2018
Accepted: 13 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
unemployment persistence; unemployment duration; non-parametric; duration dependence

HOW TO CITE:
Nonyana JZ, Njuho PM. Modelling the length of time spent in an unemployment state in South Africa. S Afr J Sci. 2018;114(11/12), Art. #4313, 7 pages. https://doi.org/10.17159/sajs.2018/4313

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
Stats SA

The deteriorating global economic conditions have worsened the unemployment situation, especially among the youth in sub-Saharan Africa. Structural factors such as the length of time spent in unemployment and job sustainability have a considerable effect on the persistence of unemployment for an individual. Non-parametric models were fitted to data consisting of 4.9 million unemployed South Africans to determine the duration dependence and probabilities associated with unemployment. The prospect of finding employment depends on unemployment duration where the rate of finding employment decreases as the length of time in unemployment increases. On average, unemployment exit is observed at lower rates, which translates to people remaining unemployed for longer durations. The human capital of the unemployed deteriorates when more time is spent in an unemployment state, thus making one less employable. Based on the Markov chain processes results, the created jobs are less sustainable because the employed transition back to an unemployment state over time. These findings suggest that the problem of unemployment in South Africa is multidimensional.

Significance:

- The structural factors associated with unemployment should be modelled to address the unemployment situation in South Africa.
- The probability of remaining unemployed increases as the length of stay in unemployment increases.
- The lengthy unemployment duration results from a low rate of exiting unemployment.

Introduction

Unemployment is a universal problem; however, the problem is more extreme in some economies. South Africa is amongst the economies with extreme levels of unemployment. The average world unemployment rate was 6.0% in 2013; the regions comprising the sub-Sahara and countries in the Organisation for Economic Co-operation and Development (OECD) recorded higher unemployment rates than the average world rate – 7.7% and 8.0%, respectively.[1,2] South Africa is one of the sub-Saharan countries with the highest unemployment rate (24.7%), together with Lesotho (24.7%), closely followed by Swaziland (22.5%)[1,3]; whilst Greece and Spain were the OECD countries with the highest unemployment rates, at 27.5% and 26.1%, respectively[2].

Unemployment conditions are associated with dire economic factors and structural factors. The main economic condition that relates to unemployment is slow economic growth, which slows the demand for labour. In most economies, slow growth is responsible for aggravated unemployment rates, whilst levels of employment improve with improvement in economic conditions.

In South Africa, higher unemployment rates are observed even when the economy is doing well. This observation suggests that unemployment in South Africa is more related to structural factors than economic factors. The main structural factors that are responsible for the current unemployment conditions in South Africa are technological advancements and a skills mismatch.[4] Statistics South Africa (Stats SA)[5] indicates that a large section of the labour force is unskilled (that is, they have an educational attainment of below a matric qualification) and many people have never been employed. Technological advancement directs employment growth towards highly skilled sectors[6], thus impacting negatively on the lowly skilled labour force[7]. The adjustment to new technologies by industries has resulted in decreased absorption rates among lowly skilled economically active persons.

The number of jobs in the manufacturing industry decreased from 2 million in 2009 to 1.8 million in 2015.[3] The decrease in jobs in the manufacturing industry is attributed to technology and utilisation of sophisticated equipment.[8] Manufacturing jobs in the USA decreased by 33% between 2000 and 2010, with technological advancement accounting for most of the decrease.[9]

Unemployment is reduced by increased levels of education.[10] However, unemployment among South African graduates has increased from a rate of 7.6% in 2008 to 9.9% in 2013. According to Altman[6], graduates' unemployment is associated with a qualifications mismatch. In addition to a qualification mismatch, Mok and Jiang[11] found that, in China, graduate unemployment is also influenced by massification of higher education. In Africa, the increase in higher education enrolments is said to be disproportionate to the increase in economic growth.[12] A notable increase of graduates has been observed in South Africa: the number of graduates with degrees or diplomas has increased by 21% between 2010 and 2014 (from 153 000 to 185 000).[13] During the same period, Stats SA recorded an increase of 109 000 in the number of unemployed people with a tertiary qualification.[3] According to Oluwajodu et al.[14], graduate unemployment in South Africa is rising with unemployment.

Lack of skills, industrial adjustment and unemployment duration are other structural factors that impact negatively on South African unemployment rates. However, these factors result from technological advancement and the skills mismatch. Few studies have focused on unemployment duration as an important factor that impacts on unemployment conditions. Studies conducted to examine the impact of structural factors in the labour market

have focused on membership of the workforce in a trade union, access to social security benefits, employment security, mismatch between job seekers and vacancies, minimum wage and factors which drive a wedge between consumer and producer prices.[10] In this study, we focused on unemployment duration as a structural factor that impacts on the current unemployment condition in South Africa.

Unemployment duration is defined as the length of time individuals spend unemployed. In South Africa, an individual is said to be in long-term unemployment if they are unemployed for a continuous period of 1 year or longer; those unemployed for a period of less than 1 year are considered to be in short-term unemployment.[15] The stability of any country correlates with its unemployment status.[16] It is thus necessary to critically model the available data with a view to finding workable solutions.

Objectives

The objectives of the study were to:

- Investigate the impact of unemployment duration on unemployment persistence.
- Determine the sustainability of jobs by predicting labour market movements.
- Make policy recommendations based on the findings.

Methodology

Data sources

We conducted a secondary analysis on data from Stats SA. Stats SA is a government department in South Africa and is responsible for the collection and publication of official data. Data collected by Stats SA are cleaned and weighted before they are posted on the Stats SA website (http://interactive.statssa.gov.za:8282/webview/) for public usage. Stats SA is solely responsible for ethical considerations.

We used panel data of 4.9 million unemployed people. The panel is created by spanning two cross-sectional data sets from a Quarterly Labour Force Survey (QLFS). QLFS is a household-based survey conducted on a quarterly basis and its sample is based on a stratified two-stage design.[15] Data on labour market activities of individuals aged 15 years and older who live in South Africa are collected for the QLFS.

The QLFS sample has features of a longitudinal survey, where 75% of the sample can be matched between two quarters. A panel is created by matching the overlapping sample. The QLFS panel data are designed to track movements of individuals between labour market status for subsequent quarters. Stats SA conducted a quality check for the QLFS panel where the results show that the data are fit for the desired purpose.

The different unemployment durations considered were: less than 3 months, 3 months to less than 6 months, 6 months to less than 9 months, 9 months to less than 1 year, 1 year to less than 3 years, 3 years to less than 5 years, and 5 years or more.

Statistical techniques

The choice of a statistical technique is dependent on the objectives and the nature of the data sets to be analysed. We applied non-parametric models because they are capable of handling incomplete observations or censored objects. The QLFS panel data showed these characteristics, as some people were not available for follow-up interviews. According to Jakoet[17], existence of right censored subjects complicates event analysis. Goel et al.[18] recommend Kaplan–Meier estimation as the best technique for computing a survival function in the presence of censored objects.

Kiefer[19] applied hazard function models to address problems such as censoring associated with duration data. Witchert and Wilke[20] recommend use of simple non-parametric models for administrative data, because of their limitations which include various forms of censoring. The unobserved heterogeneity in data sets is handled well by non-parametric models which lead to an understanding of the basics, and produce descriptive results.[21]

Studies on unemployment duration and the probabilities of leaving unemployment apply different types of data sets. Nickell[22] used cross-sectional data, Narendrathan and Stewart[23] used longitudinal data, Babucea and Danacica[24] used administrative data and Mussida[25] used rotating panel data. We applied panel data to identify the dynamic behaviour of the unemployed and control for omitted variables.

Kaplan–Meier estimator

The Kaplan–Meier estimator is a non-parametric estimator of a probability of remaining unemployed beyond time t (survival function).[26] In this method, individuals who left the study before they became employed (censored) during a given time are counted among those who survived (those still unemployed when the study concluded) and were not considered as at risk for the next period.[27] The Kaplan–Meier method sorts observations from shortest duration to longest duration, which allows for estimation of the probability of remaining unemployed beyond time t without making any assumption about the form of the function. The modelling of the probability of remaining unemployed is as follows:

Suppose T is the time it takes for an individual to secure employment. We define the proportion of those who found employment per given time (cumulative distribution) as:

$$F(t)=\Pr(T\leq t), \qquad \text{Equation 1}$$

such that the reverse cumulative equals the probability of remaining unemployed beyond time t, S(t), where

$$1\text{-}F(t)=S(t)=\Pr(T>t). \qquad \text{Equation 2}$$

The estimator of S(t) is

$$\hat{S}(t)=\prod_j^k\left(\frac{n_j-d_j}{n_j}\right), \qquad \text{Equation 3}$$

where n_j is the number of unemployed individuals (individuals at risk) at t_j and d_j is the number of employed individuals (number of events) at t_j, for $j = 1, 2, 3, 4, 5, 6$ and 7, representing the respective unemployment durations.

Nelson–Aalen estimator

The Nelson–Aalen estimator provides an efficient means of estimating the accumulated probability of exiting employment (cumulative hazard function).[28] The cumulative hazard function is then used to estimate the hazard function.

The hazard function h(t) for unemployment is the rate of exiting an unemployment state in an interval [t,t+h], for h very small positive number, given that one was in an unemployment state until time t. Hence for an infinitesimal Δt,

$$h(t)=\lim_{\Delta t\to 0}\frac{\Pr(t+\Delta t>t\,|\,T>t)}{\Delta t}=\frac{f(t)}{S(t)}. \qquad \text{Equation 4}$$

The cumulative hazard function, H(t), is defined as:

$$H(t)=-\int_0^t\frac{1}{S(u)}\left\{\frac{d}{du}S(u)\right\}du=-\ln\{S(t)\}. \qquad \text{Equation 5}$$

Hazard functions are estimated by applying smoothing techniques to the estimated cumulative hazard. However, the Kernel smoothing techniques for estimating hazard functions are inappropriate for the QLFS panel data. The time variable (unemployment durations) in the QLFS panel is categorical and the number of failures in each duration are mutually exclusive (different people are observed per duration). The Gaussian kernel, K(t), applied in the smoothing hazard has an exponential distribution of the form

$$K(t)=\frac{1}{2\pi}e^{\frac{t^2}{2}}. \qquad \text{Equation 6}$$

We applied the Kaplan–Meier type estimate in Collett[29] to estimate hazard functions for the QLFS panel.

The estimated rate of exiting unemployment in the interval t_j to t_{j+1} is defined as:

$$\hat{h}(t)=\frac{d_j}{n_j \tau_j}=e_{ji}, \qquad \text{Equation 7}$$

where $\tau_j = t_{j+1} - t_j$, and j = 1, 2 and 3 for employment, unemployment and inactivity, respectively.

Markov chains

Markov chain is a random process that changes with time, where the outcome of an experiment depends only on the outcome of the previous experiment.[30] It is a statistical technique which studies chance processes for which the knowledge of previous outcomes influences predictions for future experiments.

The chances of moving from unemployment to either employment or inactivity (transition probabilities) are predicted by using the following equation (matrix multiplication):

$$p_{ij}^{(n+m)}=\sum_{k=1}^{3} p_{ik}p_{kj}, \qquad \text{Equation 8}$$

where

$p_{ij}^{(n+m)}$ estimates the probability of leaving state *i* for state *j* in n+m steps, p_{ik} is the probability of being in state *i* in n steps, p_{kj} is the probability of being in state *j* in m steps, for $i = 1, 2, 3$; $j = 1, 2, 3$; $k = 1, 2, 3$. Let 1 = employed (E), 2 = unemployed (U) and 3 = inactivity (I).

Analysis based on non-parametric models

The Kaplan–Meier estimator is applied to estimate the chance that an unemployed person will remain unemployed in a particular unemployment duration. The results are presented in Table 1.

Table 1: Survivor functions and conditional survival probabilities for each unemployment duration considered

Unemployment duration (t_j)	Conditional survival probability (p_j)	Survivor function $\hat{S}(t_j)$
Less than 3 months	0.7652	0.7652
3 months to less than 6 months	0.7915	0.6057
6 months to less than 9 months	0.8754	0.5302
9 months to less than 1 year	0.8568	0.4543
1 year to less than 3 years	0.8942	0.4062
3 years to less than 5 years	0.9013	0.3661
5 years or more	0.9014	0.3300

Source: Computed using Q3:2013_Q4:2013 QLFS panel data.

Given the different durations of unemployment, a survival function and conditional survival probabilities are estimated. The conditional survival probability estimates the chance of remaining unemployed per given duration (exclusive), whilst the survival functions estimate the collective chance of those who remained unemployed beyond a given duration. Survival functions are estimated by first calculating conditional probabilities.

For example, the conditional survival probability (p_j) of those who searched for employment for a duration of less than 3 months, and the survival function ($\hat{S}(t)$) of those who searched for employment for a duration of 3 months to less than 6 months, are calculated, respectively, as follows:

$$p_j = P(T>t_j \mid T>t_{j-1})$$

$$= \frac{n_j - d_j}{n_j}$$

$$= \frac{631-148}{631}$$

$$= 0.7652$$

$$\hat{S}(t)=\prod_j^k p_j$$

$$= p_1 \times p_2$$

$$= 0.7652 \times 0.7915$$

$$= 0.6057$$

The results in Table 1 show that the conditional survival probabilities increase as the length of stay in unemployment increases. This implies that the likelihood of remaining in unemployment is high among those in long-term unemployment. On the other hand, the survival functions decrease as the length of stay in unemployment increases. The rate of decrease is higher among those who were in unemployment for up to 'less than 9 months'. This implies that people in short-term unemployment have higher unemployment exit rates.

The results in Table 1 further show that the probability of staying unemployed levels out for those in long-term unemployment; that is, the probability of staying unemployed is constant for all those who were unemployed for 1 year or more. This finding suggests that people get discouraged and stop engaging in job search activities after they have searched for a year. The results in Table 2 indicate that the number of those who transitioned to inactivity was the highest amongst those who were unemployed for 1 year or more. According to Stats SA releases, discouraged work seekers account for the third largest group among the economically inactive.[5]

Transitioning from unemployment to another labour market status

Between two quarters, an unemployed person can either transition to employment or inactivity or remain unemployed. Table 2 shows the number of people who left unemployment to either employment (d_j) or inactivity, and those who remained in unemployment.

The people who were unemployed during the first wave (Q3: 2013) amounted to 4.9 million and 17 000 of them did not know their unemployment duration. The numbers of unemployed people per unemployment duration are mutually exclusive, that is, different people are observed for different unemployment durations.

A large number of people who were unemployed in Q3: 2013 became inactive in Q4: 2013, compared to those who found employment: 921 000 became inactive and 638 000 found employment (Table 2).

The rate of unemployment exit

The Nelson–Aalen method was applied to estimate the rate of unemployment exit for the different unemployment durations (t_j). These rates were calculated using the labour market transitions in Table 2 and the results are presented in Figure 1.

Table 2: Labour market transitions for each unemployment duration considered

Unemployment duration (t_j)	Unemployed in Q3: 2013 (n_j)	Transitioned to employment in Q4: 2014 (d_j)	Transitioned to inactivity in Q4: 2014	Remained unemployed in Q4: 2014
	Thousand			
Less than 3 months	631	148	131	352
3 months to less than 6 months	340	71	44	225
6 months to less than 9 months	354	44	74	235
9 months to less than 1 year	347	50	64	233
1 year to less than 3 years	1223	129	223	871
3 years to less than 5 years	660	65	125	470
5 years or more	1308	129	254	926
Don't know	17	2	7	9
Total	**4880**	**638**	**921**	**3321**

Source: Computed using Q3:2013_Q4:2013 QLFS panel data.

Note: Employment refers to engagement in an economic activity; unemployed refers to those people who actively look for employment; inactive refers to people who are neither in employment nor unemployed. For more precise definitions of the labour market status see the Guide to Quarterly Labour Force Survey[15].

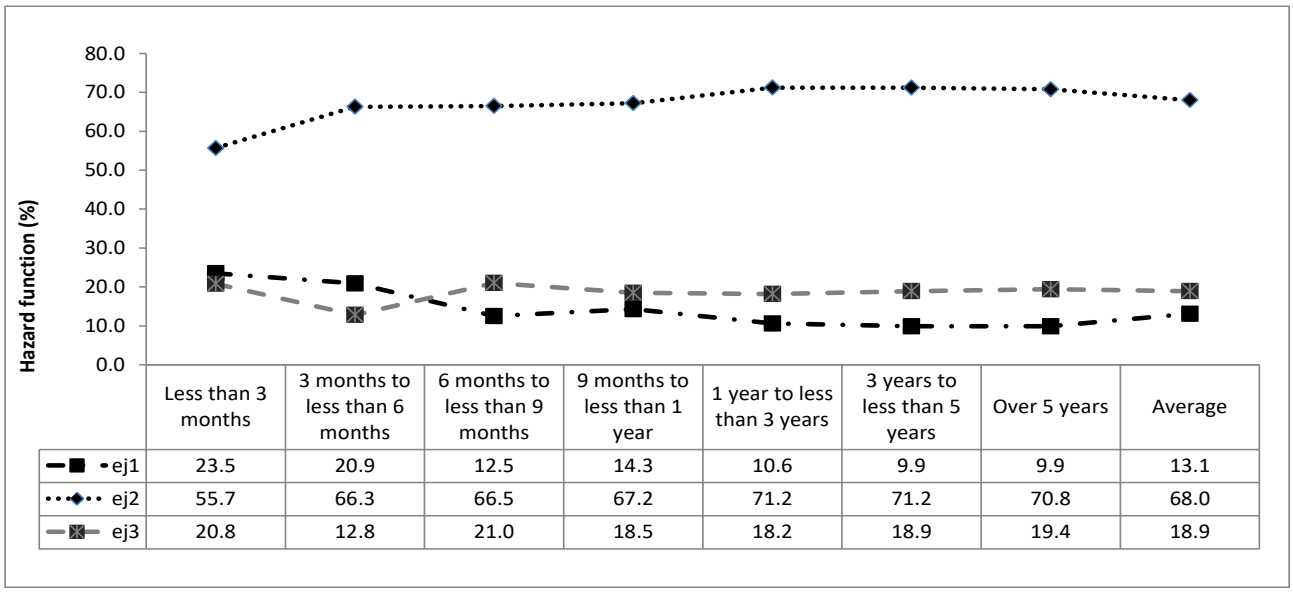


Figure 1: Hazard functions for the different unemployment durations.

For example, the estimated hazard functions (rate of finding employment -e_{11}, rate of remaining unemployed -e_{12} and rate of moving to inactivity -e_{13},) among those who looked for employment for a duration of less than 3 months are calculated using Equation 7 as follows:

$$e_{11} = \frac{d_{11}}{n_1} = \frac{148}{631} \times 100 = 23.45;$$

$$e_{12} = \frac{d_{12}}{n_1} = \frac{352}{631} \times 100 = 55.78;$$

$$e_{13} = \frac{d_{13}}{n_1} = \frac{131}{631} \times 100 = 20.76$$

Note: d_{11} is the number of people who transitioned to employment, d_{12} is the number of people who remained unemployed and d_{13} is the number of people who transitioned to inactivity.

The results in Figure 1 suggest that there are minimal employment probabilities in South Africa. On average, an unemployed person transitioned into employment at a rate of 13.1% between two quarters. In addition to the minimal employment probabilities, the likelihood of remaining unemployed is high among those who were in unemployment for a longer duration.

The findings show higher employment transitions among those with an unemployment duration of less than 6 months. However slow exit rates are observed among those who are entering the labour market for the first time (new entrants). New entrants have no work experience and have never previously sought employment.[12] The slow exit rates among

new entrants increases their stay in unemployment, which evolves into unemployment persistence.

People who were unemployed for 6 months or longer transitioned into employment at lower rates. In contrast, they left unemployment for inactivity at a higher rate. Employment transition rates for this group ranged from 9.9% (among those who were unemployed for 3 years or longer) to 14.3% (among those who were unemployed for 9 months to less than 1 year). Inactivity transition rates ranged from 18.2% (among those who were unemployed for 1 year to less than 3 years) to 21.0% (among those who were unemployed for 6 months to less than 9 months).

Duration dependence

Duration dependence measures the influence of time spent in unemployment on the rate of unemployment exit. Duration dependence is determined by calculating the rate of change as: $\Delta h(t) = \frac{dh(t)}{dt}$, assuming that h(t) is differentiable.

Unemployment duration is a categorical factor which thereby poses challenges in calculating the rate of change. The change in rate of unemployment exit $\Delta h(t)$ is used to determine duration dependence.

The prospect of getting a job is said to be dependent on time spent unemployed if the rate changes for different unemployment durations.

That is, if:

$\Delta h(t) > 0$, or

$\Delta h(t) < 0$, for all $t > 0$, then duration dependence holds.[31]

Duration dependence can be either positive or negative. Positive duration dependence happens when the rate of finding employment increases with unemployment duration. Negative duration dependence happens when the rate of finding employment decreases as unemployment duration increases. Duration dependence does not hold when the rate of finding employment remains the same between two unemployment durations.

That is, if:

$\Delta h(t) = 0$, then the rate is constant.

Table 3 indicates how the rate of unemployment exit changes as unemployment duration changes. The results in Table 3 suggest negative duration dependence, where the rate of finding employment decreased as unemployment duration increased. However, duration dependence did not hold as and when unemployment duration increased to over 5 years.

Table 3: Change in hazard functions for each unemployment duration considered

Unemployment duration (t_j)	Hazard of finding employment e_{j1}	Change in hazard functions ($\Delta h(t)$)
	%	Percentage points
Less than 3 months	23.5	
3 months to less than 6 months	20.9	-2.6
6 months to less than 9 months	12.5	-8.4
9 months to less than 1 year	14.3	1.8
1 year to less than 3 years	10.6	-3.7
3 years to less than 5 years	9.9	-0.7
5 years or more	9.9	0.0

Source: Computed using Q3:2013_Q4:2013 QLFS panel data.

Analysis based on Markov chains

We applied a Markov chain to predict transition probabilities for other quarters starting with Q1: 2014. The resulting transition matrices indicate predicted changes in labour market status (unemployment, employment and inactivity) as time increases.

The movement from one labour market state to the other defines a Markov process, and the process can start at any of the states (be it that of being employed, unemployed, or inactivity). The process started with unemployed people in Q3: 2013 who transitioned into other states in Q4: 2013.

The transition probabilities in the 3 x 3 matrix T are calculated from the QLFS panel data – Q3: 2013_Q4: 2013 and they present labour market movement between Q3: 2013 and Q4: 2013.

$$T = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}$$

$$= \begin{pmatrix} 0.929 & 0.032 & 0.039 \\ 0.131 & 0.680 & 0.189 \\ 0.041 & 0.059 & 0.900 \end{pmatrix}$$

Such that,

$P_{11} = 0.929$ is the probability of remaining employed

$P_{12} = 0.032$ is the probability of leaving employment for unemployment

$P_{13} = 0.039$ is the probability of leaving employment for inactivity

$P_{21} = 0.131$ is the probability of leaving unemployment for employment

$P_{22} = 0.680$ is the probability of remaining unemployed

$P_{23} = 0.189$ is the probability of leaving unemployment for inactivity

$P_{31} = 0.041$ is the probability of leaving inactivity for employment

$P_{32} = 0.059$ is the probability of leaving inactivity for unemployment

$P_{33} = 0.900$ is the probability of remaining inactive

In the next section, we apply a Markov chain to the 3 x 3 matrix T to predict the chances of moving from one labour market status to another (transition probabilities).

Prediction of transition probabilities

Transition probabilities for Q1: 2014 are predicted by using the matrix multiplication equation ($p_{ij}^{(n+m)} = \sum_{k=1}^{3} p_{ik} p_{kj}$).

The probability of retaining a job (p_{11}) in Q1: 2014 is estimated as follows:

$$P_{11} = P_{11}P_{11} + P_{12}P_{21} + P_{13}P_{31}$$
$$= (0.929)(0.929) + (0.032)(0.131) + (0.039)(0.041)$$
$$= 0.869$$

The other transition probabilities are calculated in the same way, and the resultant transition probability matrix for labour market movement between Q4: 2013 and Q1: 2014 is:

$$T_{Q1:2014} = \begin{pmatrix} 0.869 & 0.054 & 0.077 \\ 0.218 & 0.478 & 0.303 \\ 0.082 & 0.095 & 0.824 \end{pmatrix}$$

The matrix $T_{Q1:2014}$ shows that the probability of retaining a job between Q4: 2013 and Q1: 2014 has decreased by 6.0%, when compared to those who retained their jobs between Q3: 2013 and Q4: 2014. Of those who transitioned from employment between Q4: 2013 and Q1: 2014,

3.8% went to inactivity and 2.2% became actively engaged in job seeking activities (unemployed).

Transition probabilities for the second quarter of 2014 were predicted using the 3 × 3 matrix T and the 3 × 3 matrix $T_{Q1:2014}$, such that:

$$T_{Q2:2014} = T.T_{Q1:2014}$$

$$= \begin{pmatrix} 0.929 & 0.032 & 0.039 \\ 0.131 & 0.680 & 0.189 \\ 0.041 & 0.059 & 0.900 \end{pmatrix} \begin{pmatrix} 0.869 & 0.054 & 0.077 \\ 0.218 & 0.478 & 0.303 \\ 0.082 & 0.095 & 0.824 \end{pmatrix}$$

$$= \begin{pmatrix} 0.817 & 0.070 & 0.114 \\ 0.277 & 0.351 & 0.372 \\ 0.122 & 0.116 & 0.763 \end{pmatrix}$$

Transition probabilities for the third quarter of 2014 are predicted by solving the square of the matrix $T_{Q2:2014} = T.T_{Q1:2014}$, such that:

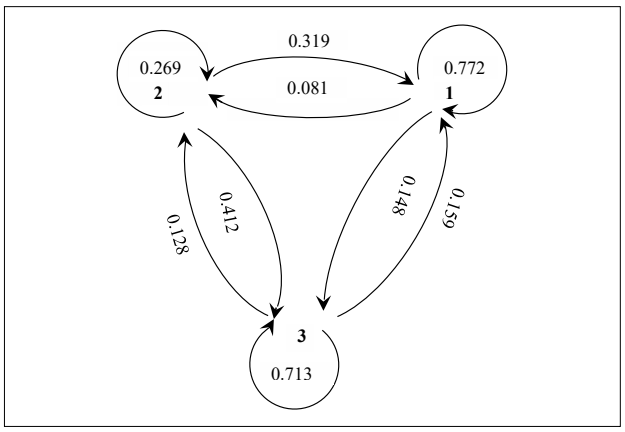
$$T_{Q3:2014} = T_{Q1:2014}.T_{Q1:2014}$$

$$= \begin{pmatrix} 0.869 & 0.054 & 0.077 \\ 0.218 & 0.478 & 0.303 \\ 0.082 & 0.095 & 0.824 \end{pmatrix} \begin{pmatrix} 0.869 & 0.054 & 0.077 \\ 0.218 & 0.478 & 0.303 \\ 0.082 & 0.095 & 0.824 \end{pmatrix}$$

$$= \begin{pmatrix} 0.772 & 0.081 & 0.148 \\ 0.319 & 0.269 & 0.412 \\ 0.159 & 0.128 & 0.713 \end{pmatrix}$$

The observed quarterly labour market movements (Q3: 2013 to Q4: 2013) as shown by matrix T present a short-term structure of the labour market. Understanding the long-term structure of a labour market is key for decision-making and planning.[32] Labour market prediction provides a basis for a long-term structure.[33]

The predicted labour market movements on matrix $T_{Q3:2014}$ are illustrated in Figure 2.



Source: Computed using Q3:2013_Q4:2013 QLFS panel data.

Note: 1 indicates the probability of remaining employed; 2 indicates the probability of remaining unemployed; 3 indicates the probability of remaining inactive.

Figure 2: Predictions of transition probabilities for the third quarter of 2014.

These predictions are for the period Q1: 2014 to Q3: 2014, where the matrix T is the input data (we have observed Q4: 2013 and we are making predictions for the three subsequent quarters). The predictions show that a person who was unemployed in Q4: 2013 had a 26.9% chance of remaining unemployed, a 31.9% chance of getting a job and a 41.2% chance of moving to inactivity in Q3: 2014. The employment retention rate is estimated at 77.2% over the period Q4: 2013 to Q3: 2014, with a 8.1% chance of moving to unemployment and a 14.8% chance of becoming inactive. Inactive people are estimated to remain in that state at a rate of 71.3%, with a 15.9% chance of moving to employment and a 12.8% chance of becoming actively involved in job search activities (unemployed).

Discussion

The findings show that the probability of leaving unemployment is not the same over a period of time. Lancaster and Nickell[34] define this character as a probabilistic process. The results in Table 1 show higher conditional probabilities which increase as the time spent unemployed lengthens. Such conditions indicate that the unemployed remained in unemployment for a long time. According to Ciuca and Matei[35], a labour market is damaging if the unemployed stay unemployed for a long time, regardless of the unemployment rate.

Figure 1 depicts higher unemployment retention rates than rates of exiting unemployment (Figure 1). The average rate of remaining unemployed is 68.0%; the lowest rate at 55.7% is for those who were unemployed for less than 3 months. In addition to the high rate of remaining unemployed, there is a greater share of those leaving unemployment for inactivity than for employment. The average rate of finding employment is 13.1%, whilst the average hazard of moving to inactivity is 18.9%. Narendranathan and Stewart[23] suggest a distinction be made between exit to employment and exit to other states.

The results indicate that unemployment exit probabilities decrease as unemployment duration increases (negative duration dependence). A Weibull analysis by Brick and Mlatsheni[36] arrived at similar results. These results suggest that the unemployed are more employable during their first 6 months in unemployment. The hazard rate of 23.5% among those who were unemployed for less than 3 months indicates limited employment opportunities.

The Markov chain processes show that the jobs created on a quarterly basis are not sustainable in the long term. While people are spending more time in unemployment, most of those who managed to exit unemployment are more likely to be without jobs within a year.

Limitations

Because survival data are characterised by censored objects and subjects have multiple entries, survival analysis techniques are therefore inadequate for analysis of mean time to failure or median time to failure. Cleves et al.[26] suggest that the point at which the survival probability is 0.5 be used as the median. It is, however, not possible to realise the point at which survival probability is exactly 0.5, because non-parametric estimates are step functions.[29]

Cleves et al.[26] recommend smoothing the discontinuities when estimating hazard functions. The standard Kernel-smoothing methodology could not be used on the QLFS panel data, because the time variable on the QLFS panel is categorical. The Kernel function applied in smoothing hazards has an exponential distribution.

Robustness of the results

The QLFS data violate the normality assumption and are also characterised by censoring. We have controlled for this challenge by using survival techniques in the analysis, as they are capable of handling censored subjects and allow the data to determine their functional form.

We applied the Kaplan–Meier type estimate in Collett[29] to estimate hazard functions for the QLFS panel, as the time variable does not meet the requirement for Kernel smoothing. Collett[29] does acknowledge that at times the use of the Kaplan–Meier type leads to irregular estimates of hazard functions. However, this method yields better results compared with the life table method because it uses exact survival times to make time stratification.

Conclusion

The South African labour market is characterised by high unemployment where the unemployed remain unemployed for longer durations. Ciuca and Matei[35] refer to such labour markets as damaging. We have found

that the average rate of exiting unemployment is very low, thus translating to lengthy unemployment duration. In addition to lower unemployment exit rates, the created jobs are less sustainable. The rate of those who transition to inactivity increases on a quarterly basis, whilst the rate of those who remain inactive over time is high. As people spend more time unemployed, their human capital and quality of life deteriorates. Without a job, people are unable to provide for the basic human needs of their families, and the level of poverty thereby increases. In trying to provide for these basic human needs, some resort to criminal activities. In addressing this social ill, we suggest a change in approach towards the unemployment problem, through development of econometric models. Effective strategies for reducing the period of unemployment can be implemented by modelling the factors that influence the period of unemployment.

Acknowledgements

The use of Stats SA data in this study is greatly acknowledged. J.Z.N. was employed by Stats SA when conducting the study and she is grateful to the management of Stats SA for instilling a culture of learning in their employees and for the financial support she received from the organisation during her studies. The University of South Africa provided a conducive environment in which to conduct this work.

Authors' contributions

J.Z.N.: Conceptualised the study; conducted the data analysis; discussed and interpreted the findings. P.M.N.: Contributed to the methodology selection, discussion and interpretation of the findings.

References

1. World Development Indicators: Unemployment. Washington DC: The World Bank Group; 2014. Available from: http://data.worldbank.org/indicator/sl.uem.totl.zs
2. Organisation for Economic Co-operation and Development (OECD). Incidence of unemployment by duration. Paris: OECD; 2014. Available from: https://stats.oecd.org/Index.aspx?DataSetCode=DUR_I
3. Statistics South Africa (Stats SA). Labour market dynamics in South Africa: Report 02-11-02. Pretoria: Stats SA; 2014.
4. Mafiri MI. Socio-economic impact of unemployment in South Africa. Pretoria: University of Pretoria; 2002.
5. Statistics South Africa (Stats SA). Quarterly labour force survey: Statistical release P0211. Pretoria: Stats SA; 2014.
6. Altman M. Youth labour market challenges in South Africa. Pretoria: HSRC; 2007. Available from: http://www.hsrc.ac.za/en/research-data/view/3620
7. Organisation for Economic Co-operation and Development (OECD). Technology, production and job creation: Best policy practices, the OECD jobs strategy. Paris: OECD; 1998. https://doi.org/10.1787/9789264163416-en
8. Danso AK. The effect of technological changes on unemployment in the beverages sector of the South African economy. Potchefstroom: North-West University; 2007.
9. Sherk J. Technology explains drop in manufacturing jobs. Washington DC: Heritage Foundation; 2010. Available from: http://report.heritage.org/bg2476
10. Morgan J, Mourougane A. What can changes in structural factors tell us about unemployment in Europe? Working paper no. 81. Frankfurt: European Central Bank; 2001.
11. Mok KH, Jiang J. Massification of higher education: Challenges for admissions and graduate employment in China. Singapore: Springer; 2017. p. 219–243.
12. Assie-Lumumba NT. Higher education in Africa: Crises, reforms and transformation. Working paper series. Dakar: CODESRIA; 2006.
13. Reddy V, Bhorat H, Powell M, Visser M, Arends A. Skills supply and demand in South Africa. Pretoria: LMIP; 2016.
14. Oluwajodu F, Blaauw D, Greyling L, Kleynhans EPJ. Graduate unemployment in South Africa: Perspectives from the banking sector. S Afr J Hum Resour Manage. 2015;13(1), Art. #656, 9 pages. https://doi.org/10.4102/sajhrm.v13i1.656
15. Statistics South Africa (Stats SA). Guide to quarterly labour force survey: Report-02-11-01. Pretoria: Stats SA; 2008.
16. Castells-Quintana D, Royuela V. Desempleo y crecimiento economico a largo plazo: el papel de la desigualdad de ingresos y la urbanizacion [Unemployment and long-run economic growth: The role of income inequality and urbanisation]. Investigaciones Regionales. 2012;24:153–173. Spanish.
17. Jakoet J. The initial unemployment duration of immigrants to Khayelitsha/Mitchell's plain. Cape Town: University of Cape Town; 2007.
18. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan–Meier estimate. Int J Ayurveda Res. 2010;1(4):274–278. https://doi.org/10.4103/0974-7788.76794
19. Kiefer NM. Economic duration data and hazard functions. J Econ Lit. 1988;26(2):464–679.
20. Witchert L, Wilke RA. Simple non-parametric estimators for unemployment duration analysis. J R Stat Soc Ser C Appl Stat. 2008;57(1):117–126. https://doi.org/10.1111/j.1467-9876.2007.00604.x
21. Mills M. Introducing survival and event history analysis. London: Sage; 2011.
22. Nickell SJ. Estimating the probability of leaving unemployment. Econometrica. 1979;47(5):1249–1266. https://doi.org/10.2307/1911961
23. Narendranathan W, Stewart MB. Modelling the probability of leaving unemployment: Competing risks models with flexible baseline hazards. J R Stat Soc Ser C Appl Stat. 1993;42(1):63–83.
24. Babucea AG, Danacica D. Using Kaplan–Meier curves for preliminary evaluation of the duration of unemployment spell. Annals of the University "Constantin Brancusi" of Targu Jiu. 2007:33–38.
25. Mussida C. Unemployment duration and compelling risks: A regional investigation. Milan: Università Cattolica del Sacro Cuore; 2007.
26. Cleves MA, Gould WW, Gutierrez RG. An introduction to survival analysis. Revised ed. College Station, TX: Taylor & Francis; 2004.
27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(283):457–481. https://doi.org/10.1080/01621459.1958.10501452
28. Klein JP, Moeschberger ML. Survival analysis techniques for censored and truncated data. 2nd ed. New York: Springer-Verlag; 2003.
29. Collett D. Modelling survival data in medical research. 2nd ed. London: Chapman and Hall; 2003.
30. Kemeny JG, Snell JL. Finite Markov chains. New York: Springer-Verlag; 1976.
31. Wooldridge JM. Econometric analysis of cross section and panel data. Illustrated reprint. Cambridge, MA: MIT Press; 2002.
32. Wilson R, Czesana V, Simova Z, Kriechel B, Vetter T. Labour market anticipation: Lessons from around the world [document on the Internet]. c2016 [cited 2018 May 19]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/608083/Labour_Market_Anticipation.pdf
33. Rihova H. Using labour market information: Guide to anticipating and matching skills and jobs. Luxembourg: European Training Foundation; 2016.
34. Lancaster T, Nickell S. The analysis of re-employment probabilities for the unemployed. J R Stat Soc Ser A. 1980;143(2):141–165. https://doi.org/10.2307/2981986
35. Ciuca V, Matei M. Survival analysis for the unemployment duration. Proceedings of the 5th WSEAS International Conference on Economy and Management Transformation; 2010 October 24–26; Timișoara, Romania. Timisoara: WSEAS Press; 2010. p. 354–359.
36. Brick K, Mlatsheni C. Examining the degree of duration dependence in the Cape Town labour market: Working paper series no.10. Cape Town: SADLRU; 2008.

Potential of marula (*Sclerocarya birrea* subsp. *caffra*) waste for the production of vinegar through surface and submerged fermentation

AUTHORS:
Tumisi B.J. Molelekoa[1]
Thierry Regnier[1]
Laura S. da Silva[1]
Wilma A. Augustyn[2]

AFFILIATIONS:
[1]Department of Biotechnology and Food Technology, Tshwane University of Technology, Pretoria, South Africa
[2]Department of Chemistry, Tshwane University of Technology, Pretoria, South Africa

CORRESPONDENCE TO:
Thierry Regnier

EMAIL:
regniert@tut.ac.za

DATES:
Received: 09 Apr. 2018
Revised: 14 Aug. 2018
Accepted: 14 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
acetic fermentation; commercialisation; fruit; high quality; sensory attributes

HOW TO CITE:
Molelekoa TBJ, Regnier T, Da Silva LS, Augustyn WA. Potential of marula (*Sclerocarya birrea* subsp. *caffra*) waste for the production of vinegar through surface and submerged fermentation. S Afr J Sci. 2018;114(11/12), Art. #4874, 6 pages. https://doi.org/10.17159/sajs.2018/4874

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
Tshwane University of Technology; National Research Foundation (South Africa)

Although there is an abundance of indigenous fruits in South Africa, knowledge of their potential uses is mainly restricted to within communities. In this study, marula fruit-processing waste by-products (fruit pulp residue and skin) were used as substrates in surface culture and submerged fermentation methods to produce vinegar (acetic acid) using spontaneous and starter culture techniques. The study revealed the possibility of producing vinegar through both methods of fermentation, with yields of acetic acid ranging between 41 000 mg/L and 57 000 mg/L (surface culture method) and between 41 000 and 54 000 mg/L (submerged culture method). Furthermore, the physicochemical property analyses revealed marula vinegar to be a potential source of bioactive compounds (total phenolics 0.289–0.356 mg/L GAE and total flavonoids 0.146–0.153 mg/L CAE) which displayed a potent antiradical activity against DPPH•: 78.85% for surface culture and 73.03% submerged culture, respectively. The sensory panel recommended application of the vinegar in products such as salad dressing and mayonnaise. Finally, we have demonstrated that the surface culture method using the inoculation technique is more suitable for the production of high-quality vinegar, with possible consideration for commercialisation.

Significance:

- Marula fruit has high economic importance for South Africa, particularly for the Limpopo Province.
- Marula waste can be a source of bioactive compounds, yet comparatively little is reported on the potential use of the waste to produce vinegar.
- Self-development of communities through viable and easy to produce commodities from marula fruit needs to be implemented and prioritised in the Limpopo Province.

Introduction

Apart from the commercial production of common fruits such as apples, peaches, pears and oranges, there is a growing trend to domesticate indigenous fruit-bearing trees in Africa such as the marula (*Sclerocarya birrea* subsp. *caffra*) and the kei apple (*Dovyalis caffra*)[1] for fruit production. Marula is particularly well known for its fruit[2], which abscise before ripening while still green and then ripen rapidly within 8 days[3]. Subsequently, the colour of the fruit changes from green to yellow, the aroma develops and the flesh softens.[3] The tree is highly appreciated by rural communities for its fruit; the edible flesh of the fruit is eaten raw or is used to prepare juices, jams, conserves, dry fruit rolls and alcoholic beverages.[4] The fruit's kernels are consumed raw or roasted, and/or used to extract oil using cold-press methods. The oil is used for cooking and is renowned for its cosmetic application.[5,6] Thus, the marula is considered a multipurpose tree in rural communities.

The resulting by-product is also further processed into value-added products. The popularity of marula is growing locally and internationally as a consequence of the well-known Amarula Cream Liqueur, manufactured locally by Distell (Stellenbosch, South Africa). Processing the fruit creates valuable waste by-products which are discarded, and as such are underutilised.

On a commercial scale, alcoholic beverage production (Amarula Cream Liqueur) and oil production for cosmetic uses are the primary commercial applications for the marula fruit. During harvest, rural communities collect the fruit and deliver it to central locations in and around the town of Phalaborwa in the Limpopo Province. Distell SA requires only 30% of the harvest for their production facility. Consequently, a significant percentage of the harvested fruit, as well as substandard fruits, are not utilised and become waste. Oelofse[7] has reported that in South Africa alone, waste generated from fruit- and vegetable-processing was 45% of various product commodities in 2012. Historically, these by-products are not considered to have commercial value because of a lack of available and affordable processes to convert the by-products into value-added commodities.

Fermentation using various microorganisms such as yeast and lactic acid bacteria is one of the oldest methods used for food preservation. Fermented foods are popular throughout the world, and make a significant contribution to the diet of millions of individuals.[8] Fermentation is a cheap and energy efficient method of preserving perishable raw materials, such as fruits and vegetables.

Vinegar is defined as a sharp sour-tasting liquid containing acetic acid obtained by fermentation of especially sour wine, malt or cider using acetic acid bacteria. It is an important condiment and typically contains ±6% (60 000 mg/L) acetic acid, carbohydrates, organic acids, alcohols and polyols, amino acids and peptides.[9]

Commercially, vinegar is produced mainly from alcoholic stock solutions such as apple ciders and grape wine, using a variety of fermentation methods. Methods include submerged and surface culture fermentation.[10] A less common approach is the use of raw agricultural crops, such as sorghum, in a solid-state fermentation type method.[11]

Vinegar production is a two-stage process, and the submerged culture fermentation method is by far the most common method in commercial production.[10] The aim of this study was to evaluate the feasibility of marula fruit waste as a substrate for vinegar production.

Materials and methods

Three batches of marula waste by-products were received from The Marula Company (Phalaborwa, Limpopo Province). The fruit waste was transported frozen in 20-L buckets to Tshwane University of Technology (Tshwane, South Africa). Upon receipt, the fruit waste was thawed in a refrigerator and working samples of 500 g were transferred into plastic Ziploc® bags, labelled MRPS1 and MRPS2 for season 1 and season 2, respectively, and stored in a freezer (Snijderg, United Scientific, Goodwood, South Africa) at -80 °C until further use.

Isolation of yeast from marula pulp

Frozen marula substrates (25 g) were subsampled from the 500 g frozen waste and defrosted under laminar flow. A 1:10 (w:v) dilution was prepared by dissolving 25 g substrate in 250 mL sterile Ringer's solution (Merck, Johannesburg, South Africa) and placed in a sterile stomacher bag and macerated with a stomacher machine (Seward, Worthing, West Sussex, England) for 5 min at 450 rpm. A 10-fold dilution series was prepared and 100 μL of each aliquot was transferred to Petri plates containing Rose Bengal Chloramphenicol Agar (Merck, Johannesburg, South Africa).

Pure yeast isolates were sub-cultured in sterile Sabouraud 2% dextrose broth. The culture was incubated at 25 °C until an optical density of 0.5 (equivalent to 1 x 10^6 cfu/mL) was obtained. Aliquot samples were taken every 24 h and optical density was measured at λ=600 nm using a spectrophotometer (Helios-Gamma, ThermoFischer Scientific, Johannesburg, South Africa). The isolates were preserved according to the method of Nyanzi[12] for further use as inoculum in alcoholic fermentation and for molecular identification. The 18S internally transcribed spacer was used to identify the yeast at species level, as recommended by Guillómon and Mas[13]. The consensus sequence was used to obtain the identity of the yeast from the UK National Centre for Biotechnology Information.

Fermentation and vinegar processing

Two fermentation methods were studied: submerged culture fermentation and surface culture fermentation. In addition, both natural fermentation and inoculated fermentation techniques were used. For the inoculation method, yeast isolated from the marula substrate was used for the alcoholic fermentation stage, and a pure culture of *Acetobacter aceti* (Anatech Cultures, Johannesburg, South Africa) was used for the acetic acid fermentation stage.

Fermentation medium

The fermentation medium was used for both methods (submerged and surface culture methods). Instant active dry yeast (*Saccharomyces cerevisiae*) (0.05 g) was added to 250 mL of warm (30±2 °C) water and left for 10 min. One tenth of the yeast was inoculated to a sterilised yeast extract peptone dextrose broth (1000 mL) and incubated at 28 °C under aeration for 24 h prior to fermentation. The vinegar processing medium consisted of 30% (w/v) marula substrate (MRPS1 or MRPS2), anhydrous glucose (in concentrations of 8%, 16% and 32% (w/v)), and yeast extract (5 mL of 2% w/v). The entire volume of each mixture was made up to 250 mL with sterile distilled water and mixed by swirling. Fermentation medium formulations yielding higher concentrations of acetic acid were considered for further physicochemical analyses.

Submerged culture fermentation

The prepared MRPS1 and MRPS2 fermentation media were inoculated with 5 mL of the naturally occurring yeast culture at an optical density of 0.5 and incubated anaerobically at 25 °C for 6 days to initiate the alcoholic fermentation stage. Subsequently, each flask was further incubated at 30 °C for 12 days and shaken at 80 rpm to aerate the medium with atmospheric air, allowing the growth of naturally occurring acetic acid bacteria (AAB) for the acetic acid fermentation stage. Fermentation was stopped by pasteurisation as follows: the resulting vinegar was aseptically transferred to a sterile 500 mL round bottom flask fitted to a Rotavapor (BÜCHI Labotechnik-AG, Flawil, Switzerland) and attached to a water bath and rotated at 80 rpm at 70 °C for 30 min. The system was closed to avoid loss of volatile compounds, and rotation ensured even distribution of heat. Subsequently the vinegar was cooled to ambient temperature and transferred into 50-mL Falcon centrifuge tubes (Eppendorf, Johannesburg, South Africa) and centrifuged at 15 810 rcf (15 °C for 10 min) (Sorvall RC 6 Centrifuge, Johannesburg, South Africa). The vinegar supernatant was collected and the pellets discarded. The vinegar was clarified by filtration (4-μm syringe filter), bottled and stored until further analyses. Analyses conducted during the fermentation process and on the final product include: pH measured on Days 6, 9, and 12 of the fermentation period, alcohol concentration measured on Day 6 using a glass alcoholmeter, and acetic acid concentration quantified using high performance liquid chromatography (HPLC) at the end of fermentation (Day 12).

Surface culture fermentation

This fermentation method followed the same procedure as stated for the submerged culture method. However, during the acetic acid fermentation stage, the flasks were not agitated, leading to the atmospheric oxygen diffusing slowly into the fermenting medium. Ethanol utilised by the AAB (in the pellicle) was quantified over time. Once depleted, fermentation was stopped by pasteurisation. The resulting vinegar was analysed as described for the submerged culture fermentation method.

Inoculated fermentation

The above-mentioned fermentation methods were mainly mediated by naturally occurring microflora, i.e. yeast and AAB. In this experiment, acetic fermentation (bioconversion of ethanol to acetic acid) was achieved by inoculation with a pure *Acetobacter aceti* (ATCC 15973) starter culture (ANATECH Cultures, Johannesburg, South Africa). Briefly, lyophilised *Acetobacter aceti* (±0.5 g) was regrown in sterile glucose yeast extract broth (250 mL) consisting of 1% (w/v) glucose, 1% (w/v) yeast extract powder, 6% (v/v) ethanol, 0.05% $MgSO_4$ and 0.05% KH_2PO_4, and incubated in a rotary incubator at 30 °C until an optical cell mass density of 0.5 was obtained. On Day 6, the fermenting medium obtained in both fermentation methods (submerged culture and surface culture methods) was inoculated with 10 mL culture broth. Alcohol, acetic acid concentration and pH were determined as previously stated.

Physicochemical analyses of the vinegar

The physicochemical analyses of the prepared vinegar solutions included: HPLC, colour assessment and determination of total phenolic content, total flavonoid content, antiradical activity and antimicrobial activity.

High-performance liquid chromatography

Organic acids (acetic, propionic and lactic acid) in the prepared vinegar solutions using both surface and submerged culture methods were quantified as described by de Sena Aquino et al.[14] Analysis of the produced organic acids was carried out through high performance liquid chromatography (Agilent Technologies 1200 Infinity, Chemetrix, Johannesburg, South Africa) equipped with an Inertsil C-18 reversed phase column (250 mm x 4.6 mm i.d. x 5 μm particle size) and a UV/Vis fixed wavelength detector at 220 nm. A mobile phase solution consisting of 0.02 mol/L KH_2PO_4 (Merck, Johannesburg, South Africa) buffer solution, adjusted to a pH (2.88 ± 0.02) was used to separate the organic acids. For calibration curves, standard solutions containing 400 μL/L, 200 μL/L, 100 μL/L and 40 μL/L of 99% organic acids (acetic, propionic and lactic) (Merck, Johannesburg, South Africa) were made up in distilled water, and the solutions filtered through 0.45 μm cellulose filter (Millipore, Johannesburg, South Africa) to remove any solid particles. The HPLC separation was performed by isocratic elution with 100% buffer at a flowrate of 1 mL/min. The total time of analysis was 10 min. The quantity of each acid present in the vinegar was determined using the following linear regression equations obtained through the standard curve calibrations:

y = 1.4045x for acetic acid;

y = 1.1563x for lactic acid and

y = 0.5237x - 12.926 for propionic acid.

Colour measurement

The colour of each vinegar at room temperature was determined using a Minolta Chromometer (Konica Minolta-CR-410, Osaka, Japan) on the basis of the CIE L*, a*, b* system.[15] Chroma and °hue angle were calculated according to Zhang et al.[16]

Total phenolic content

The concentration of phenolic compounds present in the vinegar samples was determined using the Folin–Ciocalteau method described by Du Plooy et al.[17] and expressed as gallic acid equivalent (GAE) per litre.

Total flavonoid content

The total flavonoid content of the prepared vinegars was determined following a method described by Ozturk et al.[18] and expressed as catechin acid equivalent (CAE) per litre.

Anti-radical activity

The anti-radical activity was determined as free DPPH· (2,2-diphenyl-1-picrylhydrozyl) radical scavenging capacity as described by Ozturk et al.[18], and the percentage anti-radical activity (%ARA) determined by:

$$\%ARA = \frac{Ac - As}{Ac} X100,$$

where A_c is the absorbance of the control and A_s is the absorbance of the sample.

Sensory evaluation

The sensory description of the different marula vinegars was based on a 9-point hedonic scale (1 = least like and 9 = strongly like) as described by Ubeda et al.[19] The untrained panel consisted of 23 male and female consumer science students from the Department of Hospitality Management (Tshwane University of Technology, Pretoria, South Africa) between the ages of 20 and 23 years old. White and red grape commercial vinegar samples (Wellington's®, Heinz Foods Pty Ltd, Cape Town, South Africa) were included for comparative purposes.

Statistical analysis

Each analysis was conducted in triplicate and means and standard deviations were calculated using Microsoft Excel (Microsoft Corporation, USA). The statistical significance ($p \leq 0.05$) of the data sets was evaluated using Geostats® data analysis and statistical software.

Results and discussion

Natural yeast identification

Non-*Saccharomyces* yeast (Figure 1) isolated from marula fruit by-products was identified as *Pichia kudriavzevi*. This yeast strain has been isolated worldwide from various substrates, including soil, fruits and fermentation must. This yeast species has been reported by Del Monaco et al.[20] as ideal for alcoholic fermentation in vinegar.

pH measurement

Figure 2 illustrates pH measurements for submerged and surface culture using spontaneous and inoculated bacterial cultures. In naturally fermented media, pH is controlled by the growth of a specific microorganism group (AAB or lactic acid bacteria), and yeasts by the secretion of metabolic by-products such as organic acids or organic alcohols.[13] In vinegar production, pH is the primary means of assessing the accumulation of acetic acid as a metabolic by-product in the fermentation medium.

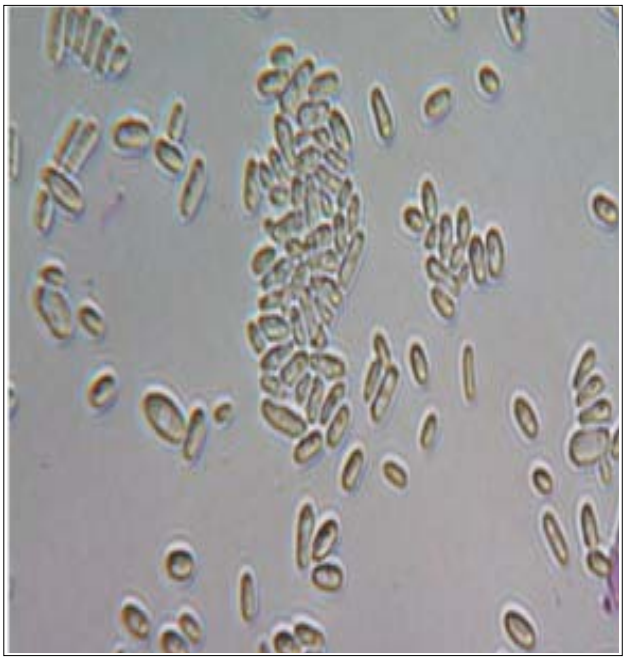


Figure 1: Naturally occurring yeast isolated from marula fruit waste by-products.

Although acetic acid is the principal component of concern in vinegar production, other organic acids such as lactic acid, propionic acid and succinic acid are also produced. Collectively, these organic acids are responsible for lowering the pH to below 6.0 in the fermenting medium. Wai-Ho et al.[21] described the optimal pH of a commercially produced vinegar to be in the range 2.0–3.5. This range is also the optimal proliferation range for AAB responsible for producing acetic acid in vinegar. The naturally fermented 8% (w/v) glucose fermentation medium had relatively lower pH values (pH 3.36–3.39) than the 16% and 32% (w/v) glucose media (pH 3.49–3.84), using both submerged and surface culture methods.

A similar trend was observed for the inoculated fermentations using 8%, 16% and 32% (w/v) glucose fermentation media. However, the inoculated fermentation method yielded lower pH (pH 2.60–3.14) for both submerged and surface culture methods. Overall, the pH values obtained in this study are higher than those reported for apple cider vinegars.[21]

It is important to remember that the functional alcohol and aldehyde dehydrogenase enzyme complexes, responsible for the bioconversion of ethanol into acetic acid, are optimally active at a pH of 2.0 to 3.5 in less than 8% alcohol (w/v).[21,22] The recorded alcohol levels for 8%, 16% and 32% (w/v) glucose media on Day 6 of alcoholic fermentation were 8%, 11% and 14% (w/v), respectively. Consequently, the higher alcohol content became inhibitory to the oxidising bacteria, resulting in vinegar with a low titre strength.

HPLC and bioactive compounds

Active compounds such as phenols and flavonoids in naturally fermented vinegars have drawn much consumer interest because of their health-conferring benefits, which include appetite suppression and free radical stabilisation. Apple cider and balsamic vinegars are by far the most renowned for these characteristics. Naturally, fermented vinegars are heterogeneous in nature, with a variety of organic acids including acetic, succinic, butyric and lactic acid.[11] This phenomenon is attributed to various pathways used by the fermentation microorganisms secreting distinct by-products and fermentation intermediates, such as acetyl aldehyde.[13]

Table 1 summarises the organic acid profile of the marula vinegar produced by the various methods. The acetic acid produced accounted for more than 99% of the organic acids. The high concentration of acetic

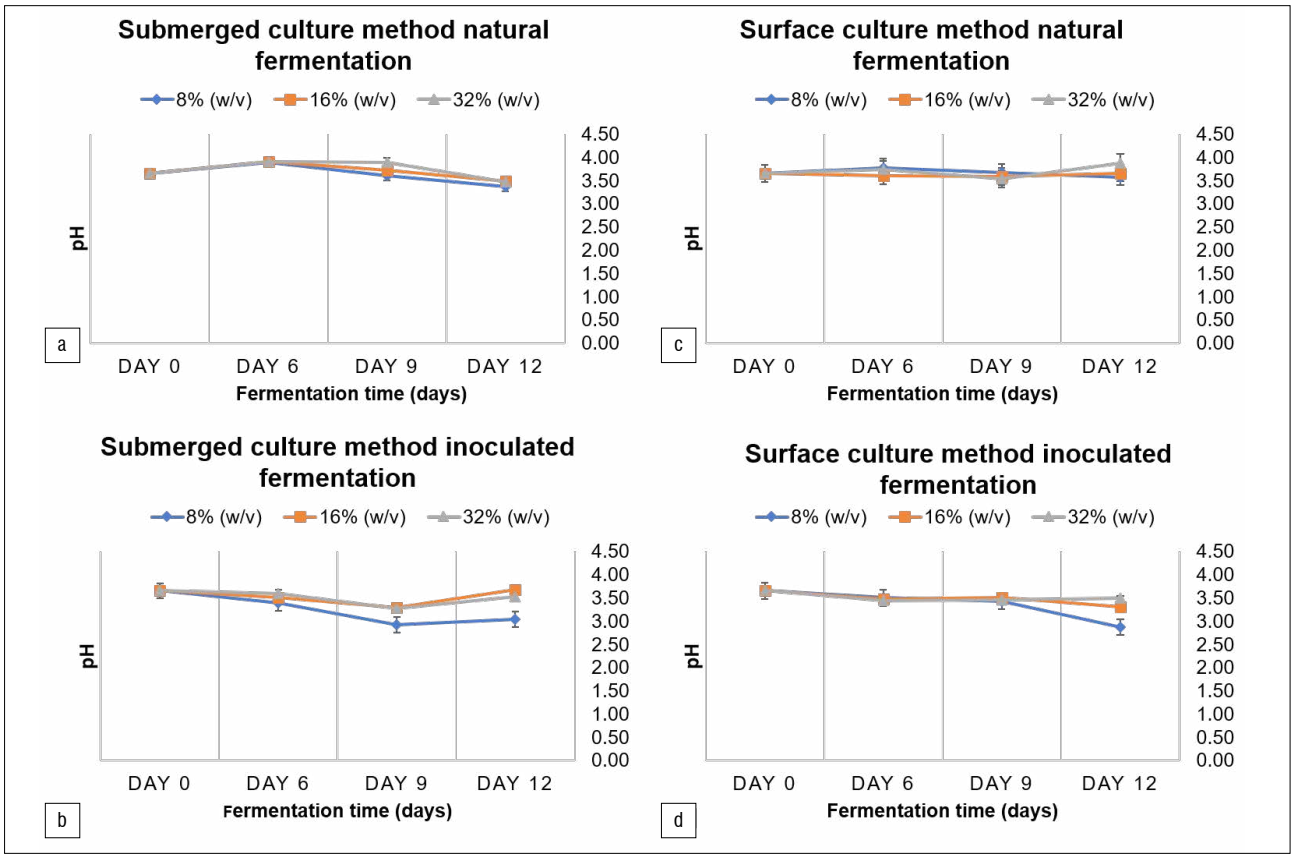


Figure 2: pH of marula vinegar over a period of 12 days, using natural and inoculated fermentation techniques, the submerged culture method (a, natural and b, inoculated) and the surface culture method (c, natural and d, inoculated). Mean values are for seasons 1 and 2 combined ($n=6$).

acid is primarily because of the dominant microorganism (AAB) present in the vinegar. These organisms are well known to produce acetic acid as a metabolic by-product.[22] In addition, more acetic acid was produced using the surface culture method combined with the inoculation technique (57 611 mg/L; Table 1) than using the submerged culture method and/or natural fermentation techniques (ranging from 41 000 to 54 000 mg/L acetic acid). The higher concentration of acetic acid produced in the marula vinegar using the surface culture fermentation method is in accordance with the findings of Tan[10], who reported that this method yields a higher quality cherry vinegar. The surface culture method is generally preferred for vinegar production, as it is non-destructive to the fermenting microorganisms, and the condition of fermentation optimises their metabolic processes.

The submerged culture method, on the other hand, requires stringent monitoring of oxygen and the replenishing of alcohol during fermentation to ensure continuous acetic acid production.[11] Moreover, the US Food and Drug Administration (FDA)[23] stipulates that the minimum threshold of acetic acid in low strength vinegars should be at least 40 000 mg/L while the Korean Ministry of Food and Drug Safety recommends a minimum of 50 000 to 85 000 mg/L acetic acid in high titre vinegars.

In the present study, the average concentration of acetic acid (ranging from 40 626 to 57 611 mg/L) in marula vinegar falls within the low strength bracket as described by the FDA.[23] Therefore, for commercialisation purposes, the optimisation of fermentation conditions (formulations and processing parameters) to achieve higher titre strength marula vinegar (≥50 000 mg/L acetic acid) will be required. This optimisation could be achieved through a rational feeding strategy of amino acids (as opposed to the addition of fermentable sugars) in the fermentation medium, as described by Zhengliang et al.[24] Nitrogenous compounds improve the enzyme complexes (ADH and ALDH) of AAB during fermentation; hence,

1. **Table 1:** Organic acid profile of marula vinegar

Organic acids	Surface culture method	Submerged culture method	Percentage difference (%)
	Natural fermentation		
Acetic acid (mg/L)	41 180.35 (± 0.086[a])	40 625.77 (± 0.172[b])	1.40
Lactic acid (mg/L)	36.09 (± 0.125[a])	19.49 (± 0.149[b])	45.99
Propionic acid (mg/L)	5.12 (± 0.170[a])	3.63 (± 0.184[b])	34.36
	Inoculated fermentation		
Acetic acid (mg/L)	57 610.54 (± 0.125[a])	53 935.09 (± 0.345[b])	6.38
Lactic acid (mg/L)	381.60 (± 0.148[a])	357.30 (± 0.125[b])	6.37
Propionic acid (mg/L)	69.96 (± 0.330[a])	88.70 (± 0.123[b])	21.13

2. *Values with the same superscripts in the same row had no significant differences ($p \leq 0.05$). Mean values are for seasons 1 and 2 combined ($n=6$).*

the bioconversion of ethanol to acetic acid by this enzyme complex becomes more efficient.

Table 2 summarises the chemical characteristics of marula vinegars produced in this study. The physicochemical properties reported in this study are in accordance with those reported by Ozturk et al.[18] for Turkish homemade vinegars. With an average phenolic content of 0.323 mg/L GAE and an anti-radical activity of ±75%, the produced vinegars contain a significant amount of secondary metabolites displaying potential anti-radical properties. These bioactive compounds are generally considered to have health benefits, as they have the ability to quench free radicals in biological systems.[25]

Table 2: Chemical characteristics of marula vinegars produced using surface and submerged culture methods with the inoculation technique

Parameters	Surface culture method	Submerged culture method
TPC (mg/L GAE)	0.356 (± 340.032[a])	0.289 (±0.023[b])
TFC (mg/L CAE)	0.153 (± 0.003[a])	0.146 (± 0.002[a])
ARA (%)	78.85 (± 0.033[a])	73.03 (± 0.033[a])

TPC, total phenolic content; TFC, total flavonoid content; ARA, anti-radical activity; GAE, gallic acid equivalent;, CAE, catechin acid equivalent

Values with the same superscript (a-b) in the same row were not significantly different (p≤0.05). Mean values are for seasons 1 and 2 combined (n=6).

Colour

Agricultural food commodities, especially fruit and vegetables, contain several colour compounds including carotene (yellow to reddish), chlorophyll (green), flavonoids (white) and anthocyanins (blue to purple).[26,27] However, adverse pH change, temperature (heat in particular), physical bruising including cutting, and processing, all influence the final colour of the destined product.[27] Marula fruit is known to contain chlorophyll and carotenes.[3] However, the colour of the marula vinegars was pale amber to deep amber. The transition from light to dark is a result of the oxidation of phenolic compounds catalysed by polyphenol oxidase. Table 3 summarises the colour values of each marula vinegar produced. It is important to notice that the fermentation methods used did not significantly change the chroma of the two vinegars. However, the marula vinegar is lighter than apple cider vinegar with an L* value of 58. The presence of negligible green (chlorophyll) colour compounds is indicated by a low negative a* value and a positive b* value. These values are in accordance with those described by Ozturk et al.[18] for Turkish homemade vinegars.

Table 3: Chroma L*, a*, and b* values for marula vinegar produced by surface and submerged culture methods

Formulation*	L*	a*	b*
	Surface culture fermentation		
Natural	58.76±0.045[a]	-1.03±0.005[c]	7.25±0.000[g]
Inoculated	57.86±0.046[a]	-1.49±0.017[d]	9.76±0.012[h]
	Submerged culture fermentation		
Natural	58.08±0.067[a]	-1.19±0.016[e]	7.46±0.029[i]
Inoculated	58.93±0.000[a]	-0.76±0.012[f]	5.63±0.009[j]

**8% (w/v) glucose fermentation medium*

Values with the same superscripts (a-j) in the same column were not significantly different (p≤0.05).

Sensory evaluation

Sensory evaluation is pivotal in the marketability of any new product. The acceptability of the new vinegars was not different from their commercial counterparts (Figure 3). While most panellists equally liked all four vinegars in term of appearance, the testers most preferred the aroma of the marula vinegar produced using the submerged fermentation method. Although the aroma of the two marula vinegars was deemed acceptable, it was described as 'uncommon' and sweet, which may be associated

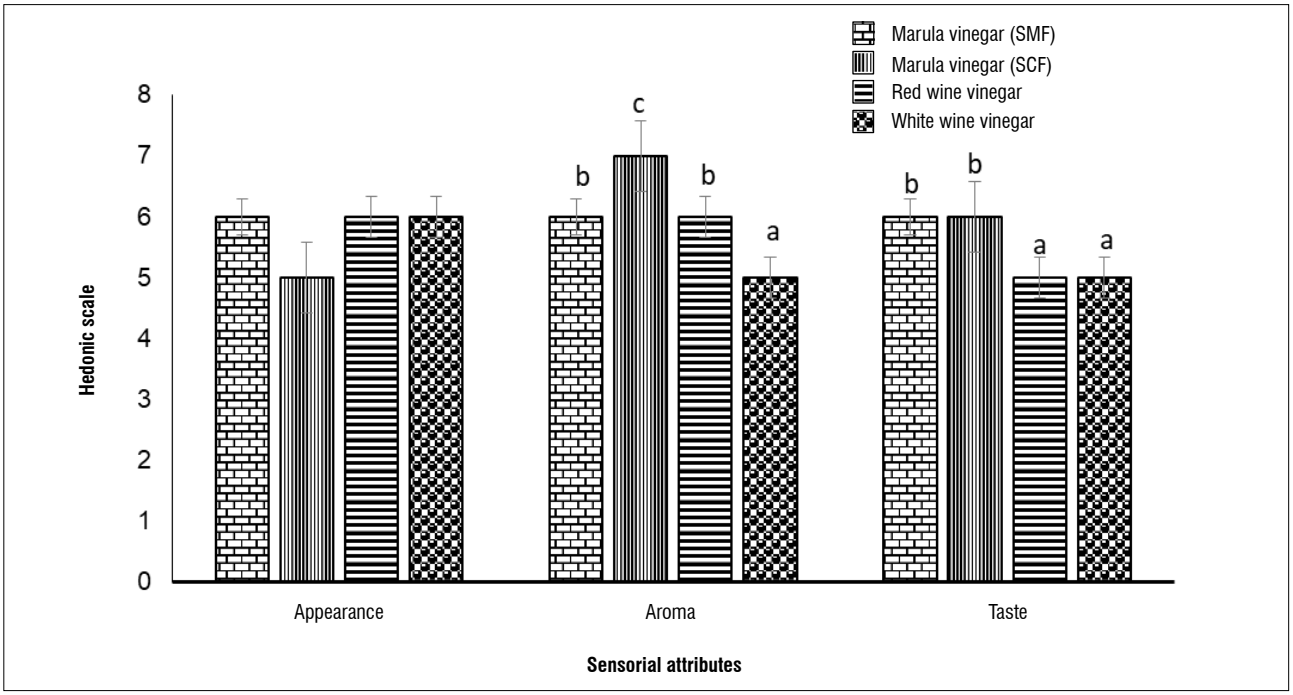


Figure 3: Sensory evaluation of marula and red and white grape vinegars using the hedonic scale. Bars with the same superscript were not significantly different ($p \leq 0.05$).

with the natural flavour of the indigenous fruit. According to Bauer et al.[28], the presence of butyl acetate can be linked to the sweet flavour of any food product. It is important to note that the taste of the marula vinegars was described as 'sweet sour' in comparison to the commercial vinegars, which were found to be 'too burny' and 'strong'. The strong taste of the commercial vinegars could be attributed to their low pH (2.57 and 2.49) and titratable acidity (7.5% acetic acid w/v) (data not illustrated) compared with the marula vinegars (pH 2.90 and 3.00, and titratable acidity 5.01% and 5.74%, respectively).

When asked about the potential use of the marula vinegars, the majority of the panellists recommended the marula vinegar to be used for the production of salad dressing or even mayonnaise. Based on the consumer's evaluation, the marula-based vinegars obtained the same consumers' purchase rating as the commercial vinegars.

Conclusion

We have shown that marula fruit-processing by-products could serve as a suitable substrate for acetic acid production. This application will add value to such products (skins, pips and pulp residues) to be utilised in tailored niche areas such as fermentation, and as such add value to the bio-economy. However, optimisation of processing (nutrient feeding strategy, temperature and aeration) parameters is necessary to ensure production of high titre marula vinegar. The surface culture method is more suitable than submerged culture fermentation for the preservation of the bioactive compounds in the vinegar. Finally, it is recommended that, with the support of the Department of Science and Technology, communities explore the production of such commodities.

Acknowledgements

We acknowledge financial support from the Tshwane University of Technology and the National Research Foundation of South Africa.

Authors' contributions

T.B.J.M. performed all the laboratory experiments as part of his master's degree and worked on the original concept of the manuscript. L.S.d.S. gave scientific inputs during the project and provided significant contributions to the final version of the manuscript. T.R. provided guidance, inputs during the research and edited the manuscript. W.A.A. provided scientific input and technical advice for the chemical analysis and edited the manuscript.

References

1. Van Wyk BE. The potential of South African plants in the development of new food and beverage products. S Afr J Bot. 2011;77:857–868. https://doi.org/10.1016/j.sajb.2011.08.003
2. Wynberg RP, Laird SA, Shackleton S, Mander C, Shackleton C, Du Plessis P, et al. Marula commercialization for sustainable and equitable livelihoods. For Trees Livelihoods. 2003;13(3):203–215. https://doi.org/10.1080/14728028.2003.9752458
3. Van Hal HP. Processing of marula (*Sclerocarya birrea* subsp. *caffra*) fruits. A case study on health promoting compounds in marula pulp (PhD thesis). Wageningen: Wageningen University; 2013.
4. Nerd A, Mizrahi Y. Domestication and introduction of marula (*Sclerocarya birrea* subsp. *caffra*) as a new crop for the Negev desert of Israel. In: Janick J, Simon JE, editors. New crops. New York: Wiley; 1993. p. 496–499.
5. Du Plessis P. Promoting indigenous fruit in Namibia. CRIAA SA-DC. Windhoek: Namibia; 2002.
6. Mojeremane W, Tshwenyane SO. The resource role of marula (*Sclerocarya birrea*): A multipurpose indigenous tree of Botswana. J Biol Sci. 2004;4:771–775. https://doi.org/10.3923/jbs.2004.771.775
7. Oleofse S. Food waste in South Africa/Africa: Opportunities and challenges. Pretoria: Council for Scientific and Industrial Research; 2013.
8. Montville TJ, Matthews KR. Food microbiology: Fundamentals and frontiers. Principles, which influence microbial growth, survival and death in foods. 2nd ed. Washington DC: ASM Press; 2001.
9. Li T, Lo YM, Moon B. Feasibility of using *Hericium erinaceus* as the substrate for vinegar fermentation. LWT – Food Sci Technol. 2014;55:323–328. https://doi.org/10.1016/j.lwt.2013.07.018
10. Tan SC. Vinegar fermentation (MSc thesis). New Orleans, LA: Louisiana State University; 2005.
11. Li S, Li P, Feng F, Xin L. Microbial diversity and their roles in the vinegar fermentation process. Appl Microbiol Biotechnol. 2015;99(12):4997–5024. https://doi.org/10.1007/s00253-015-6659-1
12. Nyanzi R. Identification and properties of potential probiotic bacteria for application in mageu (DTech thesis). Pretoria: Tshwane University of Technology; 2007.
13. Guillamòn JM, Mas A. Acetic acid bacteria. In: Carrascosa AV, Muñoz R, González R, editors. Molecular wine microbiology. New York: Elsevier; 2011. p. 227–255. https://doi.org/10.1016/B978-0-12-375021-1.10009-8
14. De Sena Aquino AC, Azevedo MS, Ribeiro DH, Costa AC, Amante ER. Validation of HPLC and CE methods for determination of organic acids in sour cassava starch wastewater. Food Chem. 2015;172:725–730. https://doi.org/10.1016/j.foodchem.2014.09.142
15. Hunterlab. Insight on color: CIE L*a*b* color scale. Reston, VA: Hunterlab; 2008.
16. Zhang Y, Wang SY, Wang CY, Zheng W. Change in strawberry phenolics, anthocyanins, and antioxidant capacity in response to high oxygen treatments. LWT – Food Sci Technol. 2007;49(1):49–57. https://doi.org/10.1016/j.lwt.2005.08.013
17. Du Plooy GW, Combrinck S, Regnier T, Botha BM. Linking lenticel discolouration of mango (*Mangifera indica* L.) fruit to reversed-phase HPLC profiles of phenolic compounds. J Hortic Sci Biotechnol. 2009;84:421–426. https://doi.org/10.1080/14620316.2009.11512543
18. Ozturk I, Caliskan O, Tornuk F, Ozcan N, Yalcin H, Balsa M, et al. Antioxidant, antimicrobial, mineral, volatile, physicochemical, and microbial characteristics of traditional home-made Turkish vinegars. LWT – Food Sci Technol. 2015;63(1):144–151. https://doi.org/10.1016/j.lwt.2015.03.003
19. Ubeda C, Callejón RM, Troncoso AM, Morales ML. Consumer acceptance of new strawberry vinegars by preference mapping. Int J Food Prop. 2017;20:2760–2771. https://doi.org/10.1080/10942912.2016.1252388
20. Del Mónaco SM, Rodríguez ME, Lopes CA. *Pichia kudriavzevii* as a representative yeast of North Patagonian wine making terroir. Int J Food Microbiol. 2016;230:31–39. https://doi.org/10.1016/j.ijfoodmicro.2016.04.017
21. Wai-Ho C, Lazim AM, Fazry S, Zaki UKH, Lim SJ. Varieties, production, composition and health benefits of vinegars: A review. Food Chem. 2017;489:1621–1630. https://doi.org/10.1016/j.foodchem.2016.10.128
22. Gullo M, Verzelloni E, Canonico M. Aerobic submerged fermentation by acetic acid bacteria for vinegar production: Process and biotechnological aspects. Process Chem. 2014;49:1571–1579. https://doi.org10.1016/j.procbio.2014.07.003
23. US Food and Drug Administration (FDA). CPG Sec. 525.825 vinegar, definitions – adulteration with vinegar eels [document on the Internet]. c1995 [cited 2018 Aug 14]. Available from: https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-afda-ice/documents/webcontent/ucm074471.pdf
24. Zhengliang-Qi Dong D, Yang H, Xia X. Improving fermented quality of cider vinegar via rational nutrient feeding strategy. Food Chem. 2017;224:312–319. https://doi.org/10.1016/j.foodchem.2016.12.078
25. Moo-Huchi VM, Estrada-Mota I, Estrada-León R, Cuevas-Glory L, Ortiz-Vázquez E, Vargas MD-LVY, et al. Determination of some physicochemical characteristics, bioactive compounds and antioxidant activity of tropical fruits from Yucatan, Mexico. Food Chem. 2014;152:508–515. https://doi.org/10.1016/j.foodchem.2013.12.013
26. Coultate TP. Food: The chemistry of its components. 5th ed. Cambridge, UK: The Royal Society of Chemistry; 2009.
27. Vagiri M, Jensen M. Influence of juice processing factors on quality of black chokeberry pomace as a future resource for colour extraction. Food Chem. 2017;217:409–417. https://doi.org/10.1016/j.foodchem.2016.08.121
28. Bauer K, Garbe D, Surburg H. Natural raw materials in the flavor and fragrance industry. In: Bauer K, Garbe D, Surburg H, editors. Common fragrance and flavor materials: Preparation, properties and uses. 3rd ed. London: WILEY-VCH; 2007. p. 161–219.

Recent emergence of CAT5 tropical cyclones in the South Indian Ocean

AUTHOR:
Jennifer Fitchett[1]

AFFILIATION:
[1]School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Jennifer Fitchett

EMAIL:
jennifer.m.fitchett@gmail.com

DATES:
Received: 12 Feb. 2018
Revised: 16 July 2018
Accepted: 21 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
hurricane; category 5; climate change; sea surface warming; southern hemisphere

HOW TO CITE:
Fitchett J. Recent emergence of CAT5 tropical cyclones in the South Indian Ocean. S Afr J Sci. 2018;114(11/12), Art. #4426, 6 pages. https://doi.org/10.17159/sajs.2018/4426

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
Society of South African Geographers

The IBTrACS global best track data set endorsed by the World Meteorological Organization provides a valuable global record of tropical cyclone genesis, track and intensity, and spans 1842 to the present. The record is significantly more robust from the late 1970s onwards, as it is supported by satellite imagery. These records indicate that the first tropical cyclone in the South Indian Ocean to intensify to CAT5 status did so in 1994. This date is significantly later than the first CAT5 storms recorded in the IBTrACS database for the Atlantic Ocean (1924) and the North Pacific (1951) recorded from ship records, and half a decade later than those of the North Indian Ocean (1989) and South Pacific (1988), captured from satellite imagery. Following this late emergence, in the period 1990–2000, eight CAT5 tropical cyclones were recorded for the South Indian Ocean. A further four have been recorded for the period 2010–2015. This recent emergence of tropical cyclones attaining category five intensity in the South Indian Ocean is of significance for the forecasting of tropical cyclone landfall and the anticipation of storm damage for the developing economies that characterise the region. Although an increase in tropical cyclone intensity is frequently projected under global climate change scenarios, the dynamics for the South Indian Ocean have remained poorly understood. Notable are early results indicating an increased frequency and poleward migration of these CAT5 storms, concurrent with a poleward migration in the position of the 26.5 °C, 28 °C and 29 °C sea surface temperature isotherms in the South Indian Ocean.

Significance:

- Category 5 tropical cyclones, the strongest category of storms, have only recently emerged in the South Indian Ocean. Since 1989, their frequency of occurrence has increased. This increase poses a heightened risk of storm damage for the South Indian Ocean Island States and the countries of the southern African subcontinent as a result of the strong winds, heavy rainfall and storm surges associated with these storms, and the large radial extent at category 5 strength.

Introduction

An increased frequency and intensity of extreme weather events is cited often as one of the most dangerous impacts projected under anthropogenic-induced climate change.[1-3] Increases in the intensity and frequency of heatwaves[4], precipitation[5] particularly in flood events[2], and tornadoes[6] have been recorded over the past five decades, and are projected to continue under even modest emission scenarios. However, despite both atmospheric and sea surface warming over the past century[7], historical analysis has demonstrated that tropical cyclone numbers have not increased for many regions in the world[8,9]. Climate models similarly project local decreases, rather than an increase in the numbers of tropical cyclones developing or making landfall in the forthcoming century.[10] This pattern is argued to be a consequence of a strengthening of atmospheric factors that inhibit tropical cyclone formation[9,10], most significant of which is an increase in vertical shear, caused by an expansion of the Hadley cell and a subsequent displacement of the subtropical jet stream[10-12]. The interaction between the warming climate and the ocean–atmospheric relationships responsible for encouraging or preventing tropical cyclone genesis lead to further debate regarding geographical changes in the occurrence of these storms. For certain ocean basins, a poleward shift in tropical cyclones, and in particular their landfall, has been observed, and a number of climate models forecast a continued poleward displacement of these systems under even the lowest carbon emission scenarios.[8,10,13] However, these results are contested by subsets of the historical record that would indicate an equatorward trajectory of the genesis latitude of these storms.[9]

Greater consensus exists regarding increasing intensities of tropical cyclones globally, both within the historical record and in modelled climate projections. This increase in storm intensity includes the maximum intensity that any given tropical cyclone will attain, the number of high intensity tropical cyclones forming (Saffir Simpson category 3–5: CAT3–CAT5), and the frequency of landfall of high intensity storms.[11,14,15] A number of studies have concluded that CAT3 and CAT4 storms, in particular, have increased in frequency globally, and are likely to continue to do so.[15] However, recent reanalyses argue that while the percentage of the most intense CAT4 and CAT5 storms are increasing in number, the frequency demonstrates a decline over the past 15 years.[16] Despite this focus on high intensity storms, the changing dynamics in the most intense storms over the past three decades remain poorly documented, amongst efforts to understand the broader scale implications of climate change on the formation of all tropical cyclones.[11]

Research on tropical cyclones in the South Indian Ocean has largely been restricted to the western half of the basin, with a particular focus on the landfall of these storms in Madagascar and Mozambique.[8,9,17-20] Many of these studies have relied on storm track data to determine spatial and temporal patterns in tropical cyclone genesis and landfall.[8,18,21] A significant focus has been placed on determining factors influencing the inter-annual variability, both in the number of storms and their geographic location within the ocean basin, highlighting the significant role of El Niño Southern Oscillation, Quasi-Biennial Oscillation, Indian Ocean Dipole and Madden–Julian Oscillation.[8,19-22] Because of the socio-economic vulnerability within the region, research on South Indian Ocean tropical cyclones

has included the considerable infrastructural and economic costs associated with storm damage.[18,23,24]

This study reflects on the emergence of tropical cyclones in the IBTrACS record for the South Indian Ocean in 1994, and the changing dynamics of these storms over the past two decades, including changes in the number of CAT5 storms, the latitudinal positions of the storms, and the underlying sea surface temperatures as drivers of cyclogenesis.

Methods

Storm track records for the South Indian Ocean were explored using the NOAA Unisys IBTrACS record (https://www.ncdc.noaa.gov/ibtracs/index.php?name=ibtracs-data). The data were explored for the period commencing on 1 June 1970 to avoid issues of data heterogeneity for earlier records which were compiled from ship logs, coastal records and a sparse distribution of aerial reconnaissance data, rather than satellite imagery used from 1970 onwards.[8,25] The compiled record terminated in 2015. The Unisys advanced filter tool was used to isolate all storms that were classified as CAT5 to determine the total decadal count of storms in each category, and the year of the first record of CAT5 storms in each ocean basin. The wind speed and central pressure of each CAT5 storm were checked in the Unisys record for the time period for which each storm was classified at CAT5 to ensure that no miscategorisation had occurred. For each storm that had attained CAT5 strength, the latitudes and longitudes of the conversion from tropical storm to CAT1 tropical cyclone, of the conversion from CAT4 to CAT5 tropical cyclone, and of the downgrade from CAT5 storm to CAT4 or lower were manually extracted together with the date of each event. The latitudes of the conversion into and dissipation from category 5 was plotted for each storm relative to the month and year of the respective events, with time-trend calculated using linear regression.

Sea surface data were obtained from the NOAA Extended Reconstructed SST V4 GrADS images (https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.ersst.html). Images were constrained geographically for the South Indian Ocean, between 0–30°S and 30–120°E, with sea surface isotherms plotted at 0.5-°C intervals. Composite plots were calculated from mean annual sea surface temperature isotherms for a long-term mean spanning 1950–2015, for decadal means for the period 1970–2009, and for the most recent half decade of tropical cyclone storm track data, 2010–2015. From these plots, the 26.5 °C and 28 °C isotherms were extracted and re-plotted to explore trends in isotherm shift of the period 1970–2015. From these plots the rate of change in the position of each isotherm was calculated at 10° longitude intervals, and averaged for the ocean basin. Weekly mean sea surface isotherm plots were obtained from NOAA V2 sea surface temperature record (https://www.esrl.noaa.gov/psd/data/gridded/data.noaa.oisst.v2.html) for the dates of each of the CAT5 tropical cyclones identified from the IBTrACS record. Two plots were generated for each time period: the first geographically constrained by the coordinate position of the storm from conversion to CAT1 through to the location of dissipation from CAT5 and the second for the entire ocean basin. From these plots, the sea surface temperature on the date and at the geographical position of the storm track on conversion to CAT1 storm, and conversion to and dissipation from CAT5 storm were extracted, together with the mean latitudinal position of the 28 °C isotherm for the South Indian Ocean basin.

Results

The first tropical cyclone classified as attaining CAT5 intensity in the South Indian Ocean is recorded in the IBTrACS record for February 1994. This date is considerably later than the first CAT5 tropical cyclones recorded in the same database for the Atlantic Ocean (1924) and the North Pacific (1951), which would have been detected from ship or coastal records. It is also half a decade later than the first CAT5 tropical cyclones recorded for the North Indian Ocean (1989) and South Pacific (1988) in the IBTrACS database. Although it could be argued that CAT5 tropical cyclones in regions of lower population density may have gone unnoticed in the early 19th century, the latter two records would have detected them by satellite imagery that provides a more globally equivalent database for all ocean basins from the 1970s onwards. While this record has been improved by direct monitoring through geo-stationary satellites[16], it is unlikely that all CAT5 storms went undetected prior to this advancement. Thereafter, four CAT5 tropical cyclones were recorded for the decadal periods 1990–1999 and 2000–2009. For the remaining 5 years captured in the data set from 2010 to 2015, a further four CAT5 storms are recorded. This recording was followed by the CAT5 tropical cyclone Fantala in April 2016.[26] A sustained occurrence of CAT5 tropical cyclones has thus been experienced following the late initiation of these very intense storms, with a slight yet statistically insignificant increase in frequency, particularly since 2010. Further monitoring of these storms should be conducted to enable a data set sufficiently large for statistical significance in trends to be determined.

The 12 CAT5 tropical cyclones recorded for the South Indian Ocean do not provide a sufficient sample size for statistically significant time trends to be detected. However, the latitudinal positions of the storm track at the time of conversion to and dissipation from CAT5 intensity both demonstrate a mean poleward trend over the period 1994–2015 (Figure 1). Notably this trend includes three of the four CAT5 tropical cyclones in the period 2010–2015 escalating to the highest strength at latitudes poleward of 16°S, and the two 2015 storms intensifying to CAT5 poleward of 17.5°S. The trend, although statistically insignificant for both intensification and dissipation, tentatively represents a poleward trajectory in CAT5 tropical cyclone intensification at a rate of 0.003°/decade (~0.33 km/decade), and dissipation at a more rapid rate of 0.004°/decade (~0.44 km/decade).

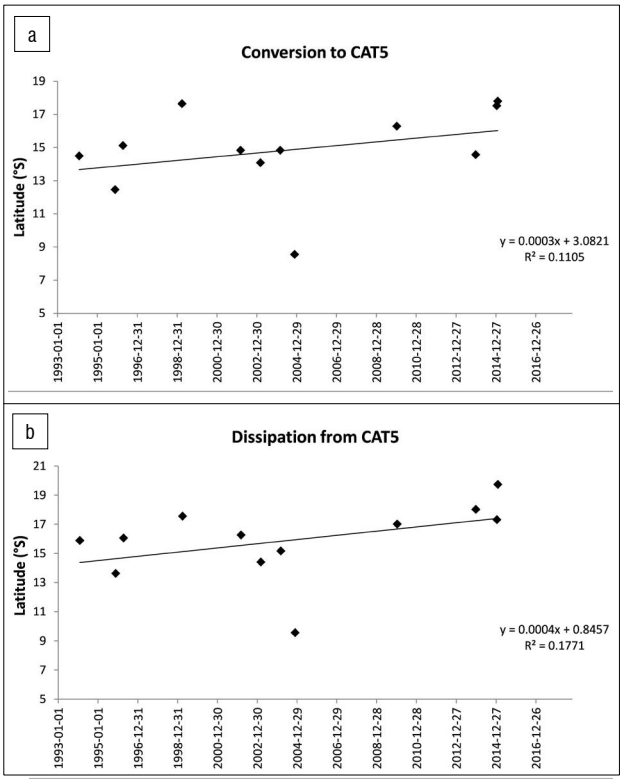


Figure 1: Latitudinal position of each CAT5 tropical cyclone storm track at the point of (a) intensification to and (b) dissipation from category 5 conditions.

The poleward shifts in the position of the storm track at the points of intensification to and dissipation from CAT5 occur concurrently with a poleward shift in the 26.5 °C sea surface temperature isotherm in the South Indian Ocean (Figure 2). This concurrence is significant because of the requirement for 26.5 °C sea surface temperatures for the cyclogenesis of tropical cyclones. The shift occurs at a mean rate of 0.068°/year (0.68°/decade) poleward for the ocean from a mean latitude of ~16°S to ~18°S over the period 1970–2015, with a more pronounced shift for the western half of the basin at a rate of 0.087°/year (0.87°/decade). For the majority of the ocean basin, this poleward

shift is progressive decade on decade. However, between 50°E and 65°E, a reversal in this trend occurs for the decade 1990–1999 with the mean 26.5 °C isotherm positioned north of the isotherm for the decade 1980–1989. The poleward shift resumes for the periods 2000–2009 and 2010–2015, but the deviation for 1990–1999 is notable given the emergence of the first CAT5 tropical cyclone in 1994 at a longitude of 58°E. The slowest latitudinal shift in the 26.5 °C is recorded for the Mozambique Channel, albeit with all isotherm positions located poleward of those for the remainder of the ocean basin. Across the ocean basin, the most rapid isotherm shift experienced was for the shortest period, from 2010 to 2015.

Although the threshold of 26.5 °C for sea surface temperatures is required for the genesis of tropical cyclones, the intensification of these storms requires additional energy supply, predominantly through latent heat. Such intensification also requires the limitation of factors that hinder cyclogenesis, most notably vertical shear. The sea surface temperature for the date and latitude of conversion from tropical storm to CAT1 tropical cyclone for all of these storms was higher than 27 °C, and was higher than 27.5°C in 8 of the 12 storms. For 9 of the 12 storms, the sea surface temperature was warmer at the point of intensification to CAT1 than at intensification to CAT5, and for all the storms, the sea surface was either the same or cooler at the point of dissipation from CAT5. A notable characteristic of all storms was that during the period of their activity, the 28 °C isotherm for the South Indian Ocean was positioned poleward of 9°S, and the 29 °C isotherm poleward of 3°S. This position is relative to a long-term 1950–2015 mean position of the 28 °C isotherm at 7°S. For the 29 °C isotherm, there is no long-term 1950–2015 mean position in the South Indian Ocean, with the first appearance of this isotherm in decadal means recorded for 1990–1999, notably coinciding with the first CAT5 tropical cyclone. Both the 28 °C and 29 °C isotherms demonstrate poleward shifts over the period 1970–2015, with the shift in the 28 °C isotherm occurring significantly more rapidly than the shift in the 26.5 °C isotherm.

Discussion

This study reveals two important findings regarding the changing characteristics of intense tropical cyclones in the South Indian Ocean. Firstly, prior to 1994, no tropical cyclones in the South Indian Ocean were classified, by the IBTrACS record, as having attained CAT5 intensity. Although it is arguable that early CAT5 tropical cyclones were not detected because of the low population density in the region and the lack of direct geostationary satellite data prior to 1989[16], post-1970 records globally have been largely standardised through satellite imagery to form the consolidated IBTrACS record. In the period 1970–1993, no tropical cyclones in the South Indian Ocean are recorded in the IBTrACS record to have intensified to CAT5, while such high intensity storms were recorded for other ocean basins. This 1994 record represents the latest global emergence of CAT5 tropical cyclones.

Secondly, between 1994 and 2015, a total of 12 CAT5 tropical cyclones are captured in the IBTrACS storm track records. Four of these storms occurred between 2010 and 2015, with a more recent CAT5 tropical cyclone in 2016 demonstrating consistency in storm intensification post-1994, and a slight, albeit statistically insignificant, increase in the frequency of CAT5 storms over time. This increase is consistent with a trend in the proportion of storms attaining CAT4 or CAT5, but at odds with the frequency of CAT4 and CAT5 storms in the South Indian Ocean.[16] This is a feature of a more rapid decline in CAT4 storms, potentially representing a greater percentage intensifying from CAT4 to CAT5 wind speed and central pressure.

Because of the low number of storms to date, the sample size is too small to detect any statistically significant trends. However, a poleward shift in the position of the storm track at intensification to and dissipation from CAT5 intensity is demonstrated through these 12 storms, concurrently with statistically significant poleward shifts in the 26.5 °C, 28 °C and 29 °C isotherms. These trends should be closely monitored to determine whether they represent persistent, long-term climatic change that supersedes inter-decadal patterns.

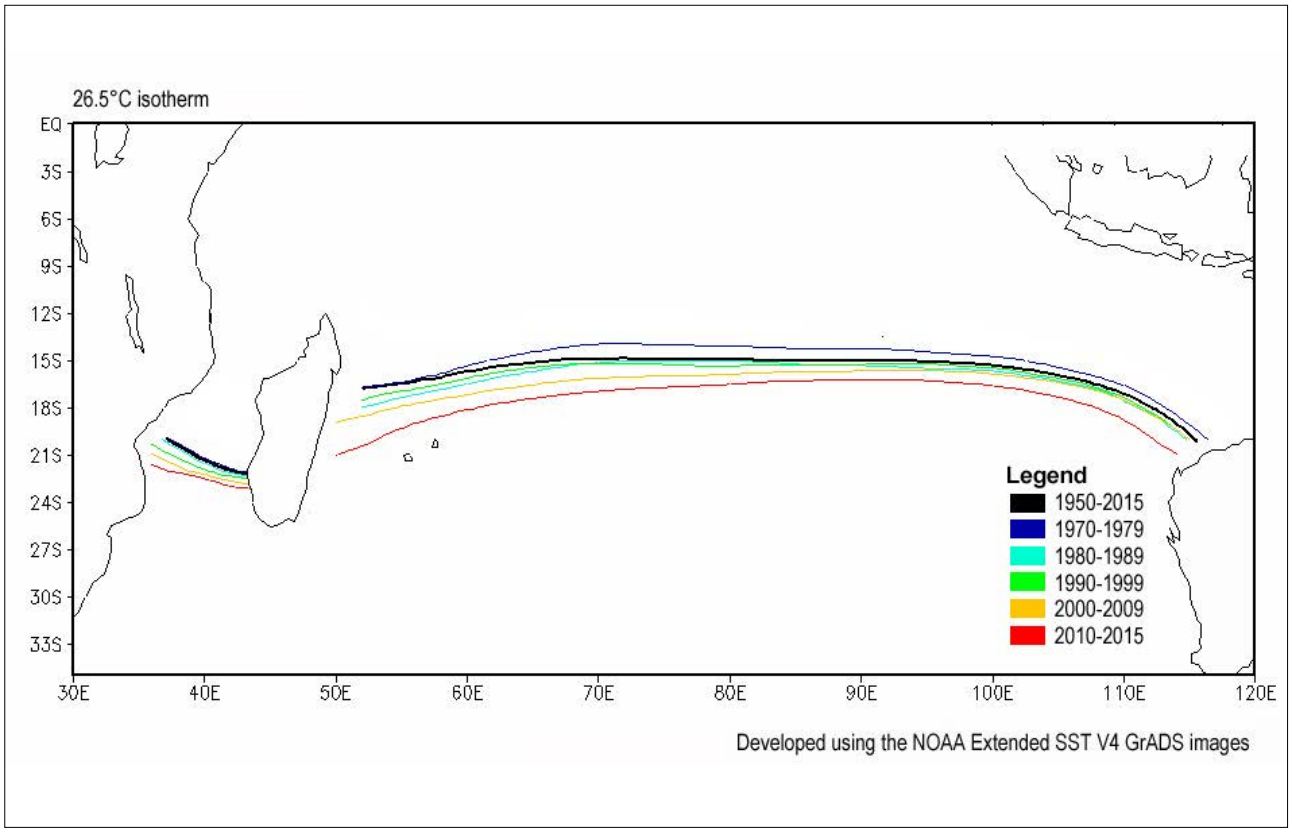


Figure 2: Decadal mean position of the 26.5 °C sea surface temperature isotherm for the South Indian Ocean over the period 1970–2015, relative to the 1950–2015 long-term mean (black line).

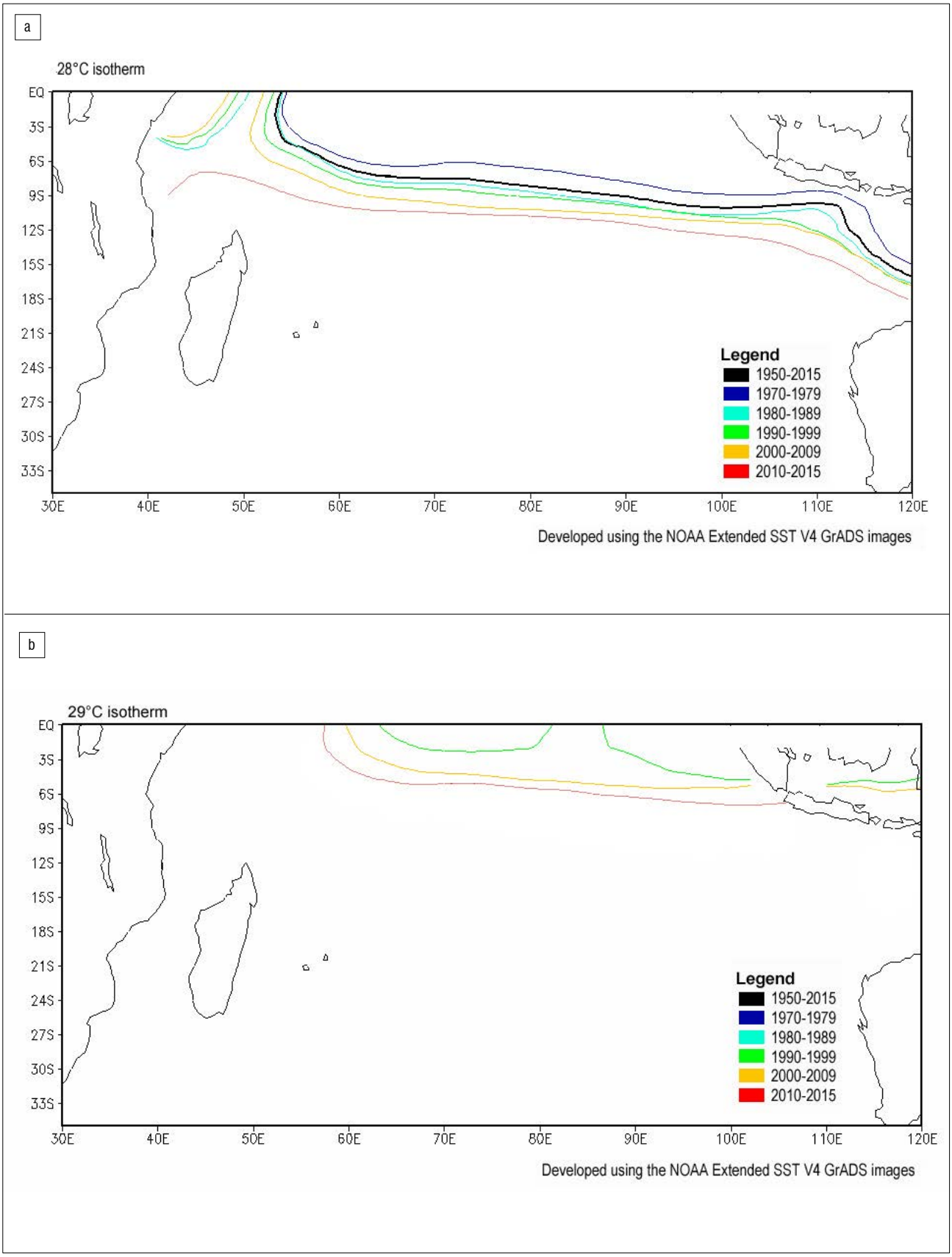


Figure 3: Decadal mean position of the (a) 28 °C and (b) 29 °C sea surface temperature isotherms for the South Indian Ocean over the period 1970–2015, relative to the 1950–2015 long-term mean (black line).

Table 1: Summary of the CAT5 tropical cyclones recorded in the IBTrACS storm track database

Year	Storm name	Date of conversion	Date of dissipation	Lowest central pressure	Highest speed
2015	Eunice	2015-01-30	2015-01-30	910 mb	140 kts
2015	Bansi	2015-01-13	2015-01-13	915 mb	140 kts
2013	Bruce	2013-12-21	2013-12-22	920 mb	140 kts
2010	Edzani	2010-01-08	2010-01-09	910 mb	140 kts
2004	Gafilo	2004-03-06	2004-03-07	895 mb	140 kts
2004	Bento	2004-11-23	2004-11-24	915 mb	140 kts
2003	Kalunde	2003-03-08	2003-03-08	910 mb	140 kts
2002	Hary	2002-03-10	2002-03-10	905 mb	140 kts
1999	Evrina	1999-03-31	1999-03-31	920 mb	140 kts
1996	Itelle	1996-04-14	1996-04-14	935 mb	140 kts
1995	Agnielle	1995-11-21	1995-11-21	927 mb	150 kts
1994	Geralda	1994-02-01	1994-02-01	905 mb	145 kts

Table 2: Sea surface temperature (SST, °C) conditions for the South Indian Ocean CAT5 tropical cyclones

Year	Storm name	SST at conversion to CAT1	SST at conversion to H5	SST at dissipation from H5
2015	Eunice	27.5	26.9	26
2015	Bansi	27.9	27.7	27.7
2013	Bruce	27	27.5	26
2010	Edzani	27	26.2	26
2004	Gafilo	28.65	28	27
2004	Bento	28.5	27.8	27.2
2003	Kalunde	29	28	27.9
2002	Hary	29	28.5	28.5
1999	Evrina	28	26.4	26.4
1996	Itelle	27	26.8	26.2
1995	Agnielle	27.1	27.5	27
1994	Geralda	27.6	28.2	28.2

The findings of this study rely heavily on the accuracy of the NOAA IBTrACS storm track database in capturing the intensity and latitudinal position of these CAT5 cyclones. These records provide an important global record of tropical cyclone activity, compiled for the period 1841–2015 through a combination of ship logs and land-based records for the early decades, and aerial reconnaissance and satellite imagery in more recent years.[25] The record includes the storm track followed by the tropical cyclone, the intensity of the storm, and the factors from which intensity is derived: central storm pressure and wind speed.[27] Concerns have been raised in the literature regarding the scientific accuracy of these records, and in particular issues of data heterogeneity as a result of the changing methods of storm identification and measurement.[16,28,29] There is some consensus that records from 1970 onwards are relatively accurate because of the use of geo-orbiting satellites that provide constant monitoring across the global oceans, rather than the more sparse coverage temporally and spatially of passing ships or individual aeroplanes on reconnaissance missions, but concern has been raised that for the Indian Ocean, direct geo-orbiting satellite data were not available prior to the launch of Meteostat-5.[11,16,27] Further efforts have been introduced to increase the accuracy of these records and their reanalysis data, with significant improvements reported[27], and the IBTrACS records are now endorsed by the World Meteorological Organization[30]. Furthermore, it is argued that the record is most accurate for storms at their maximum intensity: the storm has been in existence sufficiently long to be detected and monitored independently, and a greater impetus is placed on accurate recording as the storm develops a greater destructive potential.[11] Thus for the purpose of this study, which focuses exclusively on CAT5 tropical cyclones subsequent to 1989, the record is taken to be sufficiently reliable for calculating the findings presented.

CAT5 storms, which had not previously been recorded for the basin, occurred more frequently in the period 2010–2016 than in the previous decadal periods of 1990–1999 and 2000–2009. The Intergovernmental Panel on Climate Change has long suggested that climate change would result in an increased frequency and intensity of severe climatic events.[1] For many of the world's ocean basins, trends over the past century have demonstrated little change in tropical cyclone numbers, and for regions such as the southwest Indian Ocean and West Pacific Ocean, a decreasing trend has been observed.[8] This trend occurred concurrently with a 0.3 °C increase in mean global sea surface temperatures, which would encourage tropical cyclone formation.[7] While it has been suggested that the intensity of tropical cyclones is increasing because of this warming trend, CAT5 storm dynamics have not previously been examined in isolation.[11,15] The increase in the frequency of occurrence of CAT5 storms, particularly for a region that did not previously experience very high intensity tropical cyclones, demonstrates the manifestation of the sea surface temperature warming on tropical cyclone systems. This increase in frequency coupled with a net reduction in cumulative counts of CAT4 and CAT5 storms[16] demonstrates that the storms previously terminating at CAT4 intensity are instead increasingly reaching CAT5 intensity – a feature further reflected in the increase in the percentage of CAT4 and CAT5 storms as a subset of all tropical cyclones, both in the South Indian Ocean and globally.

The warming trend in sea surface temperatures has resulted in a poleward shift in the South Indian Ocean 26.5 °C isotherm required for tropical cyclone formation, at a statistically significant mean rate of 0.68°/decade ($p<0.0001$). More rapid is the poleward shift in the 28 °C isotherm, and the appearance and poleward migration of a 29 °C isotherm in the South Indian Ocean over recent decades. These changes are occurring concurrently with a poleward shift in the latitude of the storm track position at the points of intensification to and dissipation from CAT5 for the 12 storms on record. Moreover, a distinct relationship appears to exist between the position of the 28 °C and 29 °C sea surface temperature isotherms and the probability of occurrence of CAT5 storms, as each of the recorded storms has occurred under conditions in which the 28 °C isotherm is positioned south of 9°S latitude. These findings regarding the concurrent shifts in the 26.5 °C, 28 °C and 29 °C isotherms and the latitudinal position of CAT5 storm tracks, and the sea surface temperature distributions characterising CAT5 storms, are important for future projection of storm trajectories and intensities, and for the shorter-term forecasting during the lifespan of a storm.

These results provide a concerning outlook for the South Indian Ocean. The region comprises a number of economically developing countries and small island states, which cannot afford large capital investment in infrastructural adaptation measures to mitigate against the threats of tropical cyclones.[23,24] Unlike the United States of America, which experiences numerous tropical cyclones making landfall in any given year, only 5% of the ~9 tropical cyclones that form in the western half of the South Indian Ocean basin make landfall in any given year.[8] However, all of the storms that make landfall have devastating impacts on the livelihoods, habitat, economy and natural environment of the country affected.[24] Thus, the threat of increasing proportions of the highest

intensity tropical cyclones, even under scenarios involving an absolute reduction in storm numbers, is therefore potentially devastating. Furthermore, the poleward trajectory of these storms indicated by the 12 CAT5 tropical cyclones that have been recorded thus far, and global studies indicating a poleward trajectory in the lifetime maximum intensity of tropical storms[11], pose a heightened threat for South Africa. Although the South African coastline is currently protected from tropical cyclones by Madagascar, this southward trajectory has the potential to heighten the proportion of storms tracking south of this island nation which currently takes the brunt of tropical cyclones in the South Indian Ocean.[8,24] As one of the most economically important countries in sub-Saharan Africa, even infrequent storms pose the threat of catastrophic damage[23] for South Africa. High intensity storms would not only increase the potential for damage through the heightened wind speeds and rainfall[24], but storms of higher intensity additionally have a wider storm radius, increasing the region of damage on landfall. Considerable further monitoring of tropical cyclones in the region is warranted to confirm the trends identified from the IBTrACS storm track and GrADS V4 sea surface temperature imagery, and incorporate the findings into climate modelling efforts for the region.

Acknowledgements

This work was supported by a Society of South African Geographers Centennial Award for Emerging Career Researchers.

References

1. Intergovernmental Panel on Climate Change. Managing the risks of extreme events and disasters to advance climate change adaptation. A special report of Working Groups I and II of the Intergovernmental Panel on Climate Change. New York: Cambridge University Press; 2012.
2. Westra S, Fowler HJ, Evans JP, Alexander LV, Berg P, Johnson F, et al. Future changes to the intensity and frequency of short-duration extreme rainfall. Rev Geophys. 2014;52(3):522–555. https://doi.org/10.1002/2014RG000464
3. Fischer EM, Knutti R. Anthropogenic contribution to global occurrence of heavy-precipitation and high-temperature extremes. Nat Clim Change. 2015;5(6):560–564. https://doi.org/10.1038/nclimate2617
4. Perkins SE, Alexander LV, Nairn JR. Increasing frequency, intensity and duration of observed global heatwaves and warm spells. Geophys Res Lett. 2012;39(20), L20714, 5 pages. https://doi.org/10.1029/2012GL053361
5. Zhang X, Cong Z. Trends of precipitation intensity and frequency in hydrological regions of China from 1956 to 2005. Glob Planet Change. 2014;117:40–51. https://doi.org/10.1016/j.gloplacha.2014.03.002
6. Moore TW. A statistical analysis of the association between tropical cyclone intensity change and tornado frequency. Theor Appl Climatol. 2016;125(1–2):149–159. https://doi.org/10.1007/s00704-015-1501-3
7. Gouretski V, Kennedy J, Boyer T, Köhl T. Consistent near surface ocean warming since 1900 in two largely independent observing networks. Geophys Res Lett. 2012;39:1–8. https://doi.org/10.1029/2012GL052975
8. Fitchett JM, Grab SW. A 66-year tropical cyclone record for south-east Africa: Temporal trends in a global context. Int J Climatol. 2015;34(13):3604–3615. https://doi.org/10.1002/joc.3932
9. Malherbe J, Engelbrecht FA, Landman WA. Projected changes in tropical cyclone climatology in the Southwest Indian Ocean region under enhanced anthropogenic forcing. Clim Dynam. 2013;40:2867–2886. https://doi.org/10.1007/s00382-012-1635-2
10. Knutson TR, McBride JL, Chan J, Emanuel K, Holland G, Landsea C, et al. Tropical cyclones and climate change. Nat Geosci. 2010;3(3):157–163. https://doi.org/10.1038/ngeo779
11. Kossin JP, Emanuel KA, Vecchi GA. The poleward migration of the location of tropical cyclone maximum intensity. Nature. 2014;509(7500):349–352. https://doi.org/10.1038/nature13278
12. Henderson-Sellers A, Zhang H, Berz G, Emanuel K, Gray W, Landsea C, et al. Tropical cyclones and global climate change: A post-IPCC assessment. Bull Am Meteorol Soc. 1998;79(1):19–39. https://doi.org/10.1175/1520-0477
13. Yoshida K, Sugi M, Mizuta R, Murakami H, Ishii M. Future changes in tropical cyclone activity in high-resolution large-ensemble simulations. Geophys Res Lett. 2017;44(19):9910–9917. https://doi.org/10.1002/2017GL075058
14. Wing AA, Emanuel K, Solomon S. On the factors affecting trends and variability in tropical cyclone potential intensity. Geophys Res Lett. 2015;42(20):8669–8677. https://doi.org/10.1002/2015GL066145
15. Kuleshov Y, Fawcett R, Qi L, Trewin B, Jones D, McBride J, et al. Trends in tropical cyclones in the South Indian Ocean and the South Pacific Ocean. J Geophys Res Atmos. 2010;115(D1), Art. #D01101, 9 pages. https://doi.org/10.1029/2009JD012372
16. Klotzbach PJ, Landsea CW. Extremely intense hurricanes: Revisiting Webster et al. (2005) after 10 years. J Clim. 2014;28:7621–7629. https://doi.org/10.1175/JCLI-D-15-0188.1
17. Matyas CJ. Tropical cyclone formation and motion in the Mozambique Channel. Int J Climatol. 2015;35(3):375–390. https://doi.org/10.1002/joc.3985
18. Malherbe J, Engelbrecht FA, Landman WA, Engelbrecht CJ. Tropical systems from the southwest Indian Ocean making landfall over the Limpopo River Basin, southern Africa: A historical perspective. Int J Climatol. 2012;32(7):1018–1032. https://doi.org/10.1002/joc.2320
19. Astier N, Plu M, Claud C. Associations between tropical cyclone activity in the Southwest Indian Ocean and El Niño Southern Oscillation. Atmos Sci Lett. 2015;16(4):506–511. https://doi.org/10.1002/asl.589
20. Burns JM, Subrahmanyam B, Nyadjro ES, Murty VSN. Tropical cyclone activity over the Southwest Tropical Indian Ocean. J Geophys Res Oceans. 2016;121(8):6389–6402. https://doi.org/10.1002/2016JC011992
21. Ho CH, Kim JH, Jeong JH, Kim HS, Chen D. Variation of tropical cyclone activity in the South Indian Ocean: El Niño–Southern Oscillation and Madden–Julian Oscillation effects. J Geophys Res Atmos. 2006;111(D22), Art. #D22101, 9 pages. https://doi.org/10.1029/2006JD007289
22. Tsuboi A, Takemi T, Yoneyama K. Seasonal environmental characteristics for the tropical cyclone genesis in the Indian Ocean during the CINDY2011/DYNAMO field experiment. Atmosphere. 2016;7(5):66. https://doi.org/10.3390/atmos7050066
23. Fitchett JM, Hoogendoorn G, Swemmer AM. Economic costs of the 2012 floods on tourism in the Mopani District Municipality, South Africa. Trans R Soc S Afr. 2016;71(2):187–194. https://doi.org/10.1080/00359 9X.2016.1167788
24. Nash DJ, Pribyl K, Klein J, Endfield GH, Kniveton DR, Adamson GC. Tropical cyclone activity over Madagascar during the late nineteenth century. Int J Climatol. 2015;35(11):3249–3261. https://doi.org/10.1002/joc.4204
25. Landsea C. Counting Atlantic tropical cyclones back to 1900. EOS. 2007;88(18):197–202. https://doi.org/10.1029/2007EO180001
26. Duvat VK, Volto N, Salmon C. Impacts of category 5 tropical cyclone Fantala (April 2016) on Farquhar Atoll, Seychelles Islands, Indian Ocean. Geomorphology. 2017;298:41–62. https://doi.org/10.1016/j.geomorph.2017.09.022
27. Landsea CW, Franklin JL. Atlantic hurricane database uncertainty and presentation of a new database format. Mon Weather Rev. 2013;141(10):3576–3592. https://doi.org/10.1175/MWR-D-12-00254.1
28. Walsh KJ, McBride JL, Klotzbach PJ, Balachandran S, Camargo SJ, Holland G, et al. Tropical cyclones and climate change. WIRES Clim Change. 2016;7(1):65–89. https://doi.org/10.1002/wcc.371
29. Chan JC. Comment on "Changes in tropical cyclone number, duration, and intensity in a warming environment". Science. 2006;311(5768):1713. https://doi.org/10.1126/science.1121522
30. McTaggart-Cowan R, Galarneau Jr TJ, Bosart LF, Moore RW, Martius O. A global climatology of baroclinically influenced tropical cyclogenesis. Mon Weather Rev. 2013;141(6):1963–1989. https://doi.org/10.1175/MWR-D-12-00186.1

Applying the water-energy-food nexus to farm profitability in the Middle Breede Catchment, South Africa

AUTHORS:
Leanne Seeliger[1]
Willem P. de Clercq[1]
Willem Hoffmann[1]
James D.S. Cullis[2]
Annabel M. Horn[3]
Marlene de Witt[1]

AFFILIATIONS:
[1]Stellenbosch University Water Institute, Stellenbosch University, Stellenbosch, South Africa
[2]Aurecon Group, Cape Town, South Africa
[3]Western Cape Department of Environmental Affairs and Development Planning, Cape Town, South Africa

CORRESPONDENCE TO:
Leanne Seeliger

EMAIL:
seeliger@sun.ac.za

DATES:
Received: 07 May 2018
Revised: 21 Aug. 2018
Accepted: 22 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
water quality; farm budget models; Breede River; water security solutions; gravity-fed irrigation

HOW TO CITE:
Seeliger L, De Clercq WP, Hoffman W, Cullis JDS, Horn AM, De Witt M. Applying the water-energy-food nexus on farm profitability in the Middle Breede Catchment, South Africa. S Afr J Sci. 2018;114(11/12), Art. #5062, 10 pages. https://doi.org/10.17159/sajs.2018/5062

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
Western Cape Department of Environmental Affairs and Development Planning

The water-energy-food nexus has emerged as a useful concept to understand the multiple interdependencies that exist between the water, energy and food sectors. The nexus is an ambitious attempt to work across disciplines and scales to understand the workings of these complex systems. It is, however, criticised for being more of a general framework than a practical methodology because of the vast amount of data it would need to make real-life contributions to sustainable development. We show how the nexus approach, when used within a farm budget model, can transform the problem focus in water governance. By changing the relationship among water, energy and food production of a farm, profitability is significantly changed. The water-energy-food nexus debate is discussed within the context of the South African water sector, particularly the Breede River Catchment. Working from within the farm budget model, we demonstrate the impact of moving from an irrigation canal system that requires electricity for pumping, to a gravity-fed piped irrigation system in the Middle Breede River. The finding is that the water-energy-food nexus has the potential to unlock groundbreaking solutions to complex problems in agricultural water management when used in appropriate modelling systems.

Significance:

- The water-energy-food nexus approach can lead to an entirely new framing of water governance problems and therefore solutions to these problems.
- The water-energy-food nexus when used in farm budget models can identify ways of altering farm profitability.
- By addressing the energy cost of farming through an irrigation pipeline system in parts of the Breede Catchment Area, farm profitability could significantly increase.
- A gravity-fed closed pipeline system in parts of the Breede River can improve water availability and reduce farm and management costs.

Introduction

The devastating impacts of the current drought conditions on agriculture in South Africa's Western Cape Province and the potential risk of climate change has brought into sharp focus the complexity of issues that can affect water security on farms in this province. In this paper, we probe the power of the water-energy-food (WEF) nexus approach to address farm profitability, with the help of a hypothetical farm budget model. Whole-farm budget models, developed using spreadsheet programs, can express complex and sophisticated calculations and relationships in a relatively simple way, thereby enabling the testing of water, food and energy choices on farms. The sophistication of budget models lies in their ability to allow for detail, adaptability and user-friendliness.[1-3] The whole-farm budget can thus explore the feasibility of reconfiguring the relationship among water, energy and food, using the farm's profitability as a measure.[4]

The WEF nexus is defined as the desire to capture multiple interdependencies across three major sectors: water, energy and food.[5] It works across disciplines and across scales and spans both state and non-state actors. It is an ambitious attempt to understand complex systems using transdisciplinary research. The ambitious nature of the nexus approach, and its use as a buzzword at many conferences, has led to criticism that it is nothing more than a slogan with little practical application in real-life contexts.[6] Some of the criticism is that it needs extensive data to be useful, and that it represents general frameworks rather than useable tools.[7]

In this paper, we demonstrate how the nexus approach, when embedded in a farm budget model, can make a notable contribution to understanding how a changed relationship among water, energy and food affects farm profitability. We start with a positioning of the broad WEF nexus debate within the South African water sector and then draw links between water, energy and agricultural food production within the Breede River Catchment. In an effort to move beyond the nexus as a general framework, we calculate the impact of moving from an irrigation canal system that requires electricity for pumping, to a gravity-led piped irrigation system in the Middle Breede River, by using a farm budget model approach.

Background

The South African agricultural sector contributes approximately 2% to the gross domestic product (GDP) and employs close to 900 000 people.[8] Agriculture in the Western Cape dominates much of South Africa's agricultural export production and provides high-value products such as wine grapes and fruit. This sector is a significant employer – it provides jobs for 150 000 people across the Province.[9] South African farmers are under pressure to not only sustain, but also grow this economic contribution, despite factors putting significant financial strain on their agro-businesses and threatening their long-term profitability. This aim is expressed in South Africa's National

Development Plan that earmarks agriculture as a sector that will provide new jobs as well as address land reform issues.[10] Meanwhile, farmers are struggling to maintain production with less available water, following a drought in the Western Cape[11], and rising electricity costs that make it costly to pump and irrigate[12].

The word 'nexus' – from the Latin *nectere*, 'to bind' – speaks of connections linking two or more things.[13] The concept of nexus thinking emerged on the international agenda at the World Economic Forum in 2011 in an attempt to better understand the links between the use of resources to provide basic and universal rights to food, water and energy security. It was presented to consider key issues in food, water and energy security through a sustainability lens in order to predict and protect against potential risks of future insecurity.[14] The nexus methodology was an attempt to move away from the silo thinking that lead individual sectors to seek solutions only within their terrain of influence, and to overlook solutions that might lie in a holistic complex system approach. However, this desire to advance the basic ideas of integrated water resources management, within the water-energy-food nexus approach is not new.

Integrated Water Resources Management (IWRM) has been the dominant water management paradigm since the 1990s; it emerged from the recognition of the dysfunctional operations of sectoral approaches to water management.[15] IWRM aimed to stop fragmentary approaches to water management and high-handed development decisions made for the benefit of a single user group or faction.[16] The Global Water Partnership[17] defines IWRM as:

> *…a process that promotes the co-ordinated development and management of water, land and related resources, in order to maximise the resultant economic and social welfare in an equitable manner without compromising the sustainability of vital ecosystems.*

South Africa has been a forerunner in this international trend, with IWRM principles already emerging in the 1998 *National Water Act*[18] and the idea of integrated water management being instrumental in bringing about the idea of the Catchment Management Agencies in the country. However, despite this initial interest, it has not necessarily led to the envisaged integration of water management across sectors.[15] Water governance remains a separately managed sector. It could be argued that the emergence of the WEF nexus represents an important shift in the debate around IWRM – a shift towards a less water-centric approach to water governance. While IWRM and the nexus approach both seek to integrate water security with other policy sectors, nexus sees the food, energy and water sectors as equal partners, whereas IWRM prioritises water management.[19] We demonstrate here that this shift is important, because some water problems, such as those for which the cause lies in factors external to the water sector, cannot be addressed by a 'water-centric' approach alone.[20] In such cases, a WEF nexus approach is more effective at locating the cause of a water problem and identifying solutions to the challenge.

The value of the WEF nexus approach has already been recognised by key players in the South African water governance arena with the Water Research Commission identifying it as a new research field within water research.[21] Similarly, the World Wide Fund for Nature recognises the nexus concept as a useful framework for action to resolve complex challenges.[22] One of their key concerns has been the 20% increase in the agricultural sector's use of electricity since 2009/2010, especially when considered with the energy price hikes that have been experienced by the country over the past decade. Farmers are not easily able to pass this hike on because of the other increasing input costs in the agri-food value chain. These rising energy prices, coupled with the higher wages farmers are obliged to pay by law, are impacting the returns on investment at farm level, putting farmers – particularly emerging, previously disadvantaged farmers – under considerable pressure.

Drawing water-energy-food links in the Breede River Catchment

Agricultural food production at risk

The Breede River Catchment is located in the Western Cape Province of South Africa and is home to approximately 300 000 people.[23] The largest town in the catchment is Worcester with a population of approximately 100 000 in 2011.[23] Agriculture is the dominant economic driver and largest water user in the catchment, accounting for approximately 87% of the average annual water demand. Municipal use accounts for only 5% of the total demand, with less than 1% rural use and the remaining 7% exported to other catchments.[24]

One of the most urgent crises in the agriculture sector is the inability of farmers to pay labourers a living wage. It is estimated that agriculture and agricultural processing is responsible for 18% of jobs in the Western Cape.[25] The De Doorns farmworker strike of 2012/2013 highlighted the demand from farmworkers for higher wages. Many farmers in the Western Cape were paying above the minimum wage at the time of the strike and they claimed that they were unable to afford the ZAR150 a day that labour unions were demanding. Farmworker wages in the Western Cape are significantly higher than those in the Eastern Cape and KwaZulu-Natal Provinces.[26]

After the demonstrations, agricultural experts were concerned that farmers would be reducing labour and possibly increasing the use of machinery to reduce financial risk and save on costs in the long term. It was also forecast that some labourers would be paid better wages, but that better pay was likely to lead to job losses overall.[23] Notwithstanding these concerns, it is also clear that farm labourers are struggling to survive on the minimum wage of ZAR1503.90 a month[27] because of the high cost of food and other commodities. This situation places farming in an unstable social context, and flags the issue that unless farmers can farm more cost-effectively in the future, there is likely to be increasing social unrest in the farming sector.

Such social unrest in the Western Cape has serious consequences for South Africa as a whole, because the Western Cape produces more than 90% of South Africa's agricultural exports.[28] There is also evidence that these exports have been increasing with an annual average export growth of 17.8% between 2012 and 2016 and exports totalling ZAR121 billion in 2016 – an increase of 3.8% from 2015. However, agricultural economists see the export increase as being very strongly influenced by a depreciating rand. Moreover, while a weaker rand increases Western Cape fruit sales in Europe, it also increases the cost of imported machinery and pushes up bank interest rates. Many farmers rely on financing from banks and agricultural companies, especially during droughts when harvests are not bountiful, and for this reason the weakening of the rand puts a lot of financial pressure on the farmers.[29] This in itself could lead to further job losses and food insecurity.

Electricity cost increases

An important economic driver of change in the Breede River Catchment is the cost of electricity. A recent Water Research Commission study[30] stated that since the National Energy Regulator of South Africa in 2013 approved average annual increases in electricity tariffs of 8% for the period 2013/2014 to 2017/2018, irrigating farmers would need to re-evaluate different options to manage energy and water use in the future, because electricity costs constitute a significant part of operating costs for irrigation farmers. In 2009/2010, AgriSA reported that electricity was 4.1% of total input cost and then 5% in 2014/2015.[31] They said there had been a slight drop to 4.6% in 2016/2017 but that this drop was probably as a result of mitigation measures by farmers.[31] Fruit growers in the Western Cape are unlikely to be able to severely cut electricity consumption, because they already have to maintain the cold chain in their packing and storage processes, which again relies on electricity. With consumers being under pressure, farmers are going to struggle to pass all of the increases in cost in fruit production onto the market.[31]

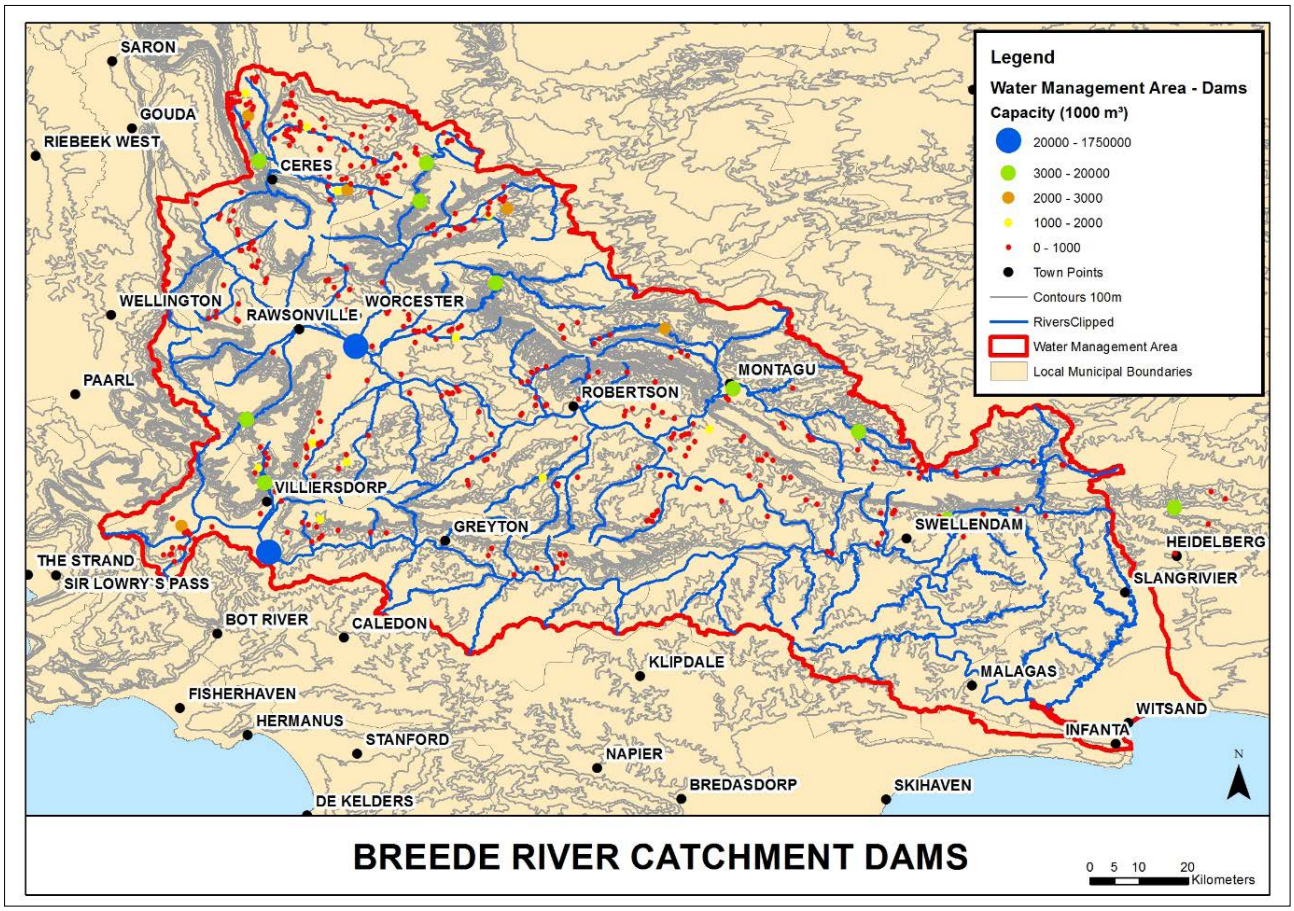
Water quantity and quality at risk

The Skurweberg Mountains near Ceres form the headwaters of the Breede River[32], which then flows in a southeasterly direction over a distance of approximately 320 km to its Indian Ocean estuary at Witsand (Figure 1). Numerous tributaries join the main stem from the mountain ranges that flank the Breede River Valley. The largest tributary of the Breede River, the Riviersonderend River, joins the Breede River approximately 20 km upstream from Swellendam. The natural mean annual run-off of the Breede–Riviersonderend system is about 1857 million m^3, and its present-day mean annual run-off is about 1156 million m^3.[24] The differences between natural and present-day mean annual run-offs do not necessarily match the sum of domestic, industrial and irrigation water requirements, because the former does not include the impacts of afforestation, invasive alien plants, irrigation/wastewater treatment return flows and nett evaporation from water bodies. In addition, a significant portion of the water (90 million m^3 per year) from the Theewaterskloof Dam on Riviersonderend River is exported to the City of Cape Town. Invasive alien plants, particularly pines and eucalypts, occurring in the upper catchment areas of the Breede River Catchment and in the riparian areas, at current levels of invasion, are estimated to account for approximately 5.3% of the registered water use, and if left unaddressed could spread to impact up to 20.7% of registered water use.[33] Because most of the Breede River Catchment receives winter rainfall and farmers harvest in summer, water users in the catchment are very dependent on dams to provide storage to meet their water use requirements. As a result, there are a large number of dams of varying sizes located within the catchment.

Intensive agriculture and urban development have resulted in an array of water quality challenges in the Breede River.[32] These challenges include concerns about increased salinity, nutrient enrichment, microbiological water quality, agrochemicals and impacts on dissolved oxygen. Water quality in the Breede River is progressively degraded in a downstream direction as a consequence of water abstractions and irrigation return flows entering the rivers. A recent study of the water quality risks in the Breede River Catchment undertaken by the Western Cape Department of Environmental Affairs and Development Planning[34] highlighted the risks associated with the maintenance and management of certain wastewater treatment works and the social and economic risks for the region associated with the state of a few of the wastewater treatment works. The potential risk to water quality posed by urban and agricultural non-point source pollutions, particularly resulting from rapid urban and informal growth, has also been described in this previous study. The economic risks associated with declining water quality in the Breede River far outweighed the costs of the necessary improvements in the condition of the wastewater treatment works, although it was also noted that some of the water-quality risks will remain because of the presence of informal settlements and continued non-point source pollution from agriculture. Eliminating the need to use the Breede River as a conveyance system for water from the main dams to the downstream farmers could therefore significantly reduce these water-quality risks for agricultural irrigation.

A farm budget model to test the water-energy-food nexus

In modern agriculture, quantitative methods are widely employed to assess the performance of industries and specific areas, within these industries, and to justify support and intervention strategies by government. Scientists use quantitative methods in innovative ways to assist food producers to make sense of complex decision-making.[35] Researchers and producers employ quantitative, scientific methods to



Produced by the Western Cape Government

Figure 1: Map of the Breede River Catchment.

communicate issues and implications to policymakers with the aim of effecting adoption.[2,36] The main advantage of simulation as a quantitative method is that the outcomes of different variables can be evaluated without actual observation and disruption within the physical system. Simulation also allows for the time-efficient and cost-saving evaluation of numerous alternative possibilities or combinations thereof.[37-39]

Farming is a complex and unpredictable endeavour. The reality of permanent crop production is that investment decisions are made without perfect knowledge of future water availability.[40] Whole-farm models are used to better understand the multiple dimensions and complex interrelationships of farms.[41,42] Modelling is an attempt to validate accurate representations of the real world. By varying the parameters and assumptions of the model, research questions of a descriptive, causal and predictive nature can be partially addressed.[43-45] Computer models are particularly useful in this regard for exploring hypothetical systems and quantitatively comparing, and designing, alternative management options in relatively stable patterns.[46-48]

Budgeting is perhaps the most widely used non-optimising method (not designed to identify the 'best' possible outcome) of financial planning. The popularity of budgets stems from their simplicity of use and the fact that they aid in the heuristic approach to decision-making.[49] The development of computer technology introduced a dimension to budgeting methods that allowed budgets to be used as dynamic planning and decision-making tools. In this sense, budgets can now also be classified as simulation models that are based on accounting principles and methods, rather than purely on mathematics.[2,50,51]

The key to useful models is relative simplicity, which can be obtained by setting well-defined objectives. The pre-occupation of systems researchers with simulation and model building, with less attention to applications, may lead to either limited practical use or suspicion among producers who do not understand the principle or function of the model.[52] Another limitation of whole-farm budget models is their non-optimising nature.

The availability of irrigation water determines, in physical terms, the crop mix and the amount of land that is useable on a farm. The physical size and geography of the farm also determines the investment required in infrastructure, movable assets and other inputs. This investment requirement, along with the harvest, determines the projected income. The income and cost factors are therefore greatly influenced by a change in irrigation water availability and quality. Net farm income is commonly used for a financial comparison of farming units. With some adaptation, whole-farm models may also be extended over time to calculate returns on capital invested and to calculate profitability indicators such as the internal rate of return on capital investment (IRR) or net present value (NPV).

What is the typical farm?

A tool that can be used to assess farm profitability and to determine the effect of variations in a range of variables on farm-level profitability is the 'typical farm'. The typical farm model allows for evaluation and comparison of the effect of various managerial decisions and options in a cost- and time-efficient way.[53] The typical or representative farm model cannot accurately reflect internal managerial problems for individual, unique farm units. The impact of trends, strategies and policy options on whole-farm profitability can be assessed by using a typical farm.[54]

The whole-farm approach provides a more reliable basis for assessing the potential impact of variables on which to base policies and programmes. The typical farm concept is based on existing enterprises, practices and environmental factors. Feuz and Skold[55] define a typical farm as 'a model farm in a frequency distribution of farms in the same universe'. In essence, it is a synthetically constructed model farm based on the expected structure of a farm in a particular area.[56]

Describing a typical farm in the Middle Breede River Catchment

The Breede River irrigation area is a diverse farming area of high-value, irrigated, mostly long-term crops. The area, however, has been subdivided into four relatively homogeneous areas. These areas are defined as relatively homogeneous based not only on soil and climate, but also on farm size, farmer association and crop type. The initial identification of the areas was based on the areas as used by the South African Wine Industry Information & Systems (SAWIS) in the farmer study group work. The farm sizes and land distribution for the typical farms was done according to statistics obtained from industry organisations.[57-59] The farm description for each area was presented to industry role players for validation of the assumptions. It was also suggested and accepted that production cost figures be used based on the industry information published by Hortgro and VinPro, representing the fruit and wine industries, respectively.

By definition, relatively homogeneous farming areas are similar in physical characteristics like climate, soil and farm size, but also in the formal and institutional-like affiliations to agribusinesses. For the Breede River, the areas that were identified are Breedekloof, Worcester, Robertson and Montagu and are shown in Figure 2. For each area, a typical farm was identified, which would be recognised as characteristic by farmers of the area. Typical is closer to modus, or the most commonly occurring, rather than average. The purpose is to establish a basis for comparing alternatives and so it should be representative. Farmers must be able to identify with the 'typical farm'.

The first dimension of whole-farm modelling is that of establishing the physical extent and processes concerned in the farm system, that is, the farm size and cultivated area. The important consideration is that not all land is included in the investment requirement: only the productive area contributes to output and income. Table 1 shows the farm size for each of the areas as well as the area that is cultivated.

The crop mix for wine grapes that is implemented on each farm is often done in conjunction with wine cellars. This crop mix is the typical cultivar mix. SAWIS and Hortgro data were used for the initial land-use pattern identification. These data were presented to industry role players for validation. The validated land-use patterns for each typical farm are presented in Table 2. This group included personnel from VinPro, Hortgro and SAWIS, as well as producers, winemakers and viticulturists from cellars; Intelegro and Nexus (chemical suppliers); and Yara and Nitrophoska (fertiliser suppliers). The industry role players validated both the land-use patterns of the typical farm and the investments, costs and yields, to establish the gross margins of the farm within each homogeneous region.

For each farm, a set of assumptions is required for quantities and prices for both inputs and products. The cost and expected income were taken from industry bodies: VinPro does financial study group work for the wine industry and Hortgro for the fruit industry. Production costs are thus based on the outcome of study group data for each area – an average cost amongst a group of producers ranging from 16 to about 26 participants. Fixed improvements were also determined with the assistance of land valuators and farmers. The same is the case for the fruit industry. Prices for land were obtained from banks and farm valuators and not from estate agents. The farm values might therefore be somewhat conservative as the farms that require financing are often in some difficulty. Profitability is a function of income generated by some investment. The investment requirement for the typical farm for each area is presented in rand value in Table 3. The investment requirement thus includes investment in land, fixed improvements (orchards, vineyards, buildings, houses, irrigation dams and pump stations and fencing) and movable assets (tractors, implements, equipment).

The operational component of a farm is expressed in the model in the form of an enterprise budget. The enterprise budgets were compiled for each wine cultivar separately and integrated into the whole-farm model by multiplying the area under each crop by the expected profitability. Table 4 shows the enterprise budget for the Chenin Blanc wine cultivar for Breedekloof as example of the structure. The model is developed for a 25-year period, but only the first 6 years are presented in Table 4, from Year 6 onwards the vineyard will expectedly run on full bearing capacity until the last year. The model is parameterised to the extent that input and output quantities can be changed separately. This functionality allows for the model to be used for risk assessment and the identification of key drivers of profitability.

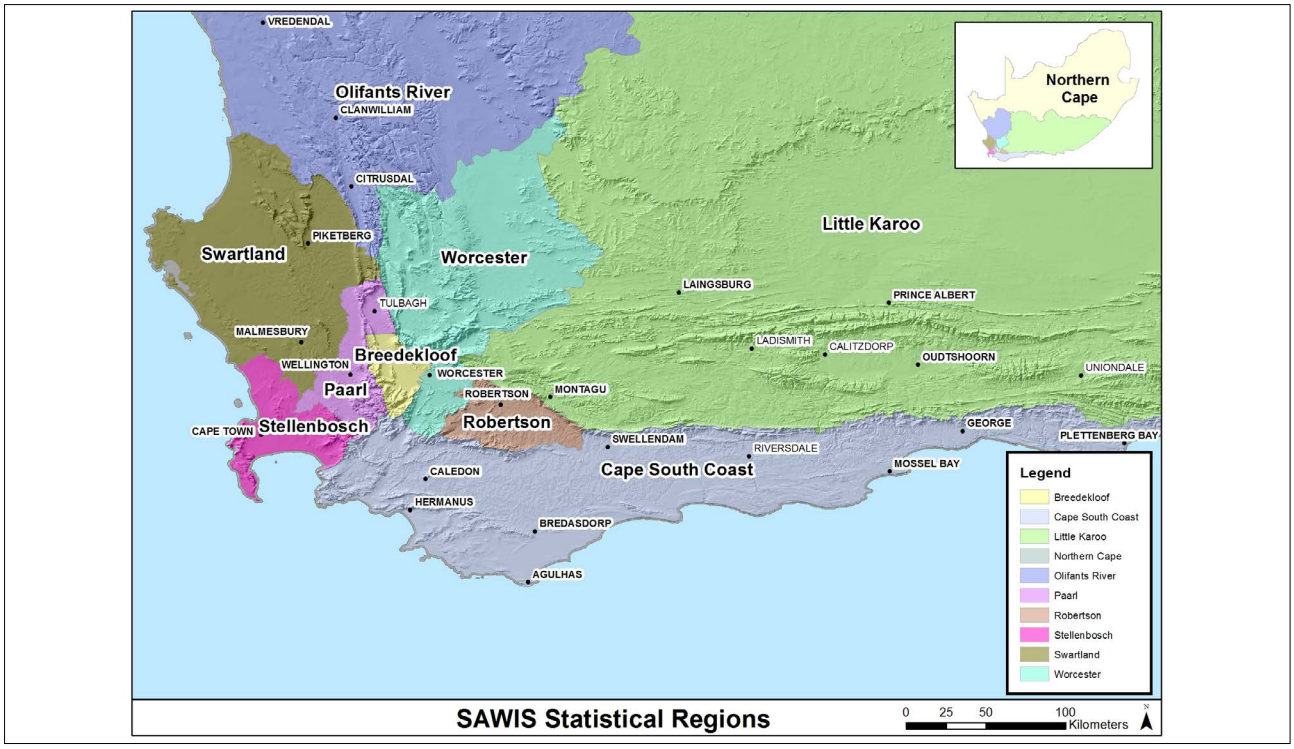


Figure 2: The wine production regions as defined by the South African Wine Industry Information & Systems[60].

Table 1: Farm size and cultivated area for each relatively homogeneous area in the Breede River area

Area	Farm size (hectares)	Cultivated area (hectares)†
Breedekloof	90	73
Worcester	140	42
Robertson	200	76
Montagu	270	48

†*Non-cultivated areas could be roads, housing and infrastructure, dams, windrows, salinised areas, or simply lack of irrigation water.*

Table 2: Land-use system (ha) for the 'typical farm' for each relatively homogeneous area

	Breedekloof	Worcester	Robertson	Montagu
Wine grapes				
Chenin Blanc	14.6	8.8	9.0	12.2
Colombar	12.4	7.5	11.6	18.4
Sauvignon Blanc	5.8	3.5	9.0	5.7
Chardonnay	8.8	5.3	10.3	0.4
Cabernet Sauvignon	8.0	4.9	8.4	0.0
Pinotage	12.4	7.5	7.1	0.0
Shiraz	7.3	4.4	5.8	4.1
Merlot	3.7	2.2	3.2	0.0
Ruby Cabernet	3.7	2.2	3.9	8.2
Fruit				
Peaches				
Keisie		1.2	1.4	0.7
Kakemas		0.7	1.4	0.7
Oom Sarel		1.4	1.0	0
Sandvliet		0.3	0.3	0
Neethlings		0.2	0.2	0
Malherbes		0.3	0.4	0
Cascade		0.7	0.9	0
Plums				
Souvenir			1.1	0
Harry Pickstone			1.7	0
Apricots				
Bulida			2.9	5.8

Table 3: The investment requirement in rands (ZAR) for the 'typical farm' for each area

Item / area	Breedekloof	Worcester	Robertson	Montagu
Land	22 477 050	18 561 700	25 930 100	13 662 000
Fixed improvements	9 791 333	9 144 083	9 519 583	9 197 983
Movable assets	3 722 345	4 348 345	5 267 345	4 320 754
Total	35 990 728	32 054 128	40 717 028	27 180 737

The gross production value is a function of yield and expected price. The yield assumptions are relatively conservative, and as shown in Table 4, are expected to reach full bearing capacity in Year 6. The establishment cost occurs in Year 1 of the life of a vineyard. From Year 2 onwards, the production costs will steadily increase according to the yield expectations. The gross margin serves as a basis for the modelling component and it is the difference between gross production value and allocated variable cost. For this cultivar the expected gross margin stabilises at ZAR50 198 for the remainder of the expected life of the vineyard over 20 years.

The various enterprises are integrated into the whole-farm gross margin level, by subtracting the variable costs from the income. Fixed and overhead costs are then subtracted from the gross margin to calculate a figure that would resemble net farm income. From this net farm income annual figure, the capital replacement is subtracted to give the net annual flow after capital replacement. The model is structured to show the impact of changes in: input or output quantities, input and output prices, fixed costs levels, changes in land utilisation, crop replacement schedules and movable asset replacement costs. Profitability is an indication of yield on investment and in this case is indicated by the IRR and NPV. These profitability indicators are calculated on a capital budget format and thus include capital replacements in normal farming cycles (machinery and orchards and vineyards). The normal increase in land values is ignored. It is important to note that the IRR in this case is a real return, in other words one should still add inflation to calculate the nominal rate of return. Table 5 shows the expected IRR and NPV for each typical farm.

Table 4: The enterprise budget for Chenin Blanc for Breedekloof

Chenin Blanc						
Year	**1**	**2**	**3**	**4**	**5**	**6**
Income						
Expected yield	0%	0%	12%	31%	71%	**100%**
Gross production value	**0**	**0**	**8 413**	**21 735**	**49 780**	**70 112**
Directly allocated variable cost						
Establishment cost (total ZAR)	**171 628**	**3382**	**0**	**0**	**0**	**0**
Land preparation	30 750	0	0	0	0	0
Drainage system	5519	0	0	0	0	0
Trellis system	69 302	0	0	0	0	0
Plant material	42 272	3382	0	0	0	0
Irrigation system	23 785	0	0	0	0	0
Pre-harvest cost (total ZAR)	**3735**	**3735**	**5254**	**7659**	**12 723**	**16 394**
Fertilisers	0	0	218	562	1287	1813
Irrigation	3735	3735	3735	3735	3735	3735
Insecticides	0	0	288	744	1704	2400
Fungicides	0	0	114	295	675	950
Herbicides	0	0	115	298	682	960
Labour	0	0	168	433	991	1396
Fuel	0	0	287	742	1698	2392
Repairs and maintenance	0	0	330	852	1951	2748
Harvest cost (total ZAR)	**0**	**0**	**324**	**837**	**1917**	**2700**
Labour	0	0	324	837	1917	2700
Other (total ZAR)	**8768**	**356**	**263**	**383**	**636**	**820**
Sundry	8768	356	263	383	636	820
Total	**184 131**	**7473**	**5841**	**8879**	**15 276**	**19 914**
Gross margin (ZAR)	**184 131**	**7473**	**2573**	**12 855**	**34 503**	**50 198**

Table 5: The expected internal rate of return on capital investment (IRR) and net present value (NPV) for the 'typical farm' of each area

	IRR (%)	NPV (ZAR)
Breedekloof	3.25%	3 453 796
Worcester	3.35%	3 310 039
Robertson	4.43%	9 793 890
Montagu	4.12%	5 008 049

Reconfiguring the water-energy-food nexus

As discussed earlier, water quality and quantity are both serious threats to water security in the Breede River Catchment. It is predicted that the Western Cape will possibly become drier and drier over the years as climate change predictions become a reality.[61] Many of the farms in the Breede River Catchment receive water via canal systems, which at some points are over 100 years old. Figure 3 illustrates the canal systems in the catchment.

Large amounts of the already little available surface water disappear through leakages and evaporation in this inefficient system. While many farms have access to groundwater, some of this groundwater is brackish because of the geology of the area and return flows. The middle and lower reaches of the Breede River are most affected, with salinity levels progressively increasing in a downstream direction.[63] These high salinity levels force farmers to irrigate more, to leach out the salts. Indications are that farmers are over-irrigating to compensate for salinity and other forms of pollution. Approximately 22 million m^3/a of fresh water is released from the Greater Brandvlei Dam into the Breede River to reduce salinity of the river water.

Use of a gravity-led piped system from the dams would improve water quality by supplying clean water fresh from the dams and not allowing the geological salinity of the Breede River Catchment soils, nor the agri-polluted return flows from the farmers or the storm water from informal settlements, to contaminate river water for irrigation, and subsequently the soil. This improvement in the water quality would reduce over irrigation of soils to produce good food, resulting in less water having to be used by farmers in the entire system. The freshwater releases from the Greater Brandvlei Dam, which are currently used to dilute the pollutants in the water, could therefore be minimised.

Applying the reconfigured nexus approach to the farm budget model

The farm budget model can provide an understanding of the feasibility of replacing the current open canal system with a gravity-led piped system. The expected effect of irrigation and electricity is measured against the baseline presented in Table 5. Three scenarios were modelled. In the first scenario, more and less irrigated land is made available, which reflects the possibility of increasing or reducing the water available for irrigation, and the impact that this has on farm profitability was assessed. Both 2 ha and 4 ha increases in land use were simulated in each case. In the second scenario, it was assumed that water is delivered under pressure through a pipeline system with various reductions in electricity as a cost, together with the maintenance and cost of pumps. This assumption has a direct effect on cost saving and profitability. A third scenario was based on the possibility of more water being made available for crop production, with a specific application to a crop of white grapes. Producers may increase irrigation quantity for white wine cultivars and stimulate increased yields (personal communications with researchers and producers). The increase in water quantity leads to a 5% yield increase of white grape cultivars, with differing profitability in the different homogeneous areas. This scenario is very much dependent on

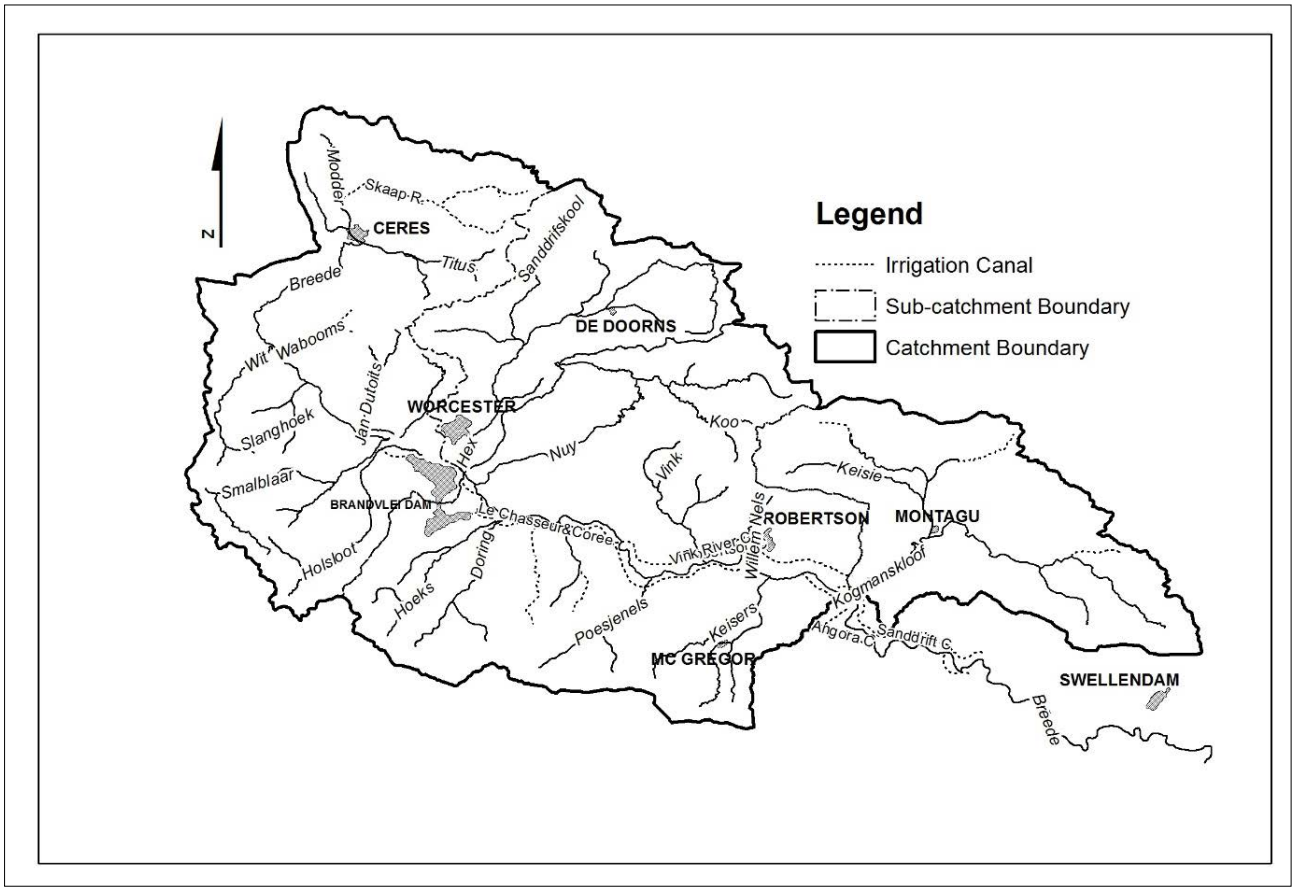


Figure adapted from Kienzle and Flügel[62].

Figure 3: Map showing the irrigation canal system of the Upper and Middle Breede River.

individual cellars and their target markets. The correlation between high yields and lower quality wine may exclude this option in some areas, although cellars may target lower price or wine for distillation markets. The expected financial outcome for each of the scenarios is presented in Table 6.

The analysis shows a high positive impact on IRR and NPV in all regions under Scenario 2, where a gravity-fed piped system cuts electricity costs, with 30%, 60% and 100% reductions which are based on the topographical fall of the land in relation to the source of the water. The model shows that, given the high cost of electricity, with further price increases expected, by the provision of gravity-fed piped water, which provides the pressure for irrigation water, and the subsequent reduction in electricity, the profitability of the typical farm can increase. With a 5–10% increase in irrigated land (approximately 4 ha) across all typical farms in this analysis in the Breede River Catchment, as in Scenario 1, there are strong parallels to a 30% reduction in electricity in Scenario 2. This analysis proves that placing emphasis on the energy element of the water-energy-food nexus in a catchment management system can have extremely positive impacts on food production and economic growth at the farm level.

The idea of introducing solar energy powered pumps was not included in the analysis. It is worth investigating as a supplementary form of energy. It is, however, unlikely to be suitable as an exclusive form of energy generation because farmers in this region pump both during the day and night. This means if solar energy was introduced, farmers would need to use batteries to store the solar energy. This would involve a significant capital outlay and would need to be the subject of another research exercise.

Farmers' opinions on the impacts of the piped system

Farmers and irrigation board members interviewed in the Cogmanskloof Area near Montagu (Figure 1) were in favour of using a piped gravity-fed system in their area, given that the altitude drops from the catchment dam to their farms. They assessed that 25% of the water was lost through evaporation and leaks in the aging canal system. An estimated ZAR13 million investment in 20 km of gravity-fed piped water from Potjieskloofdam would significantly reduce their electricity costs and save them sufficient quantities to significantly increase the profitability of their farms. The irrigation board projected that the irrigation system for the area would become so profitable, the board would be able to release water for about 50 ha to the government for distribution to emerging farmers, subject to more detailed analysis by water specialists. It was estimated that repairing the current canal system that was ageing would be equally as expensive as implementing the new pipeline system, but that the repairs would make no contribution to improving the water, food and energy crisis in the Montagu area. However, in the lower Robertson area with water travelling significant distances from the higher catchment to reach smaller incremental altitudinal decreases onto a flatter plateau, stakeholders were less confident that a gravity-fed irrigation system could be put in place. They argued that the capital cost of piping water from Brandvlei Dam to the farms in the Breede River lower down was too high and the gradient insufficient to significantly reduce the use of electrical pumps. The hypothetical testing of the value of a reconfiguration of the WEF nexus in the Middle Breede Area using the farm budget model enables the cost of electricity to be adjusted to gauge projected impact of the gravity-fed piped irrigation system.

Conclusion

It was not the aim of this paper to provide conclusive evidence of the use of a piped-irrigation system versus a canal system in the Breede River Catchment, but rather to demonstrate how the adoption of a WEF nexus approach for agricultural water management in the Breede allowed a new problem identification to emerge. By focusing on the relationship between water and energy, rather than water scarcity alone, more water was made available at a reduced cost and with improved quality in the Middle Breede River Catchment. Moreover, if addressed, the analysis showed that it would not only increase farm profitability and address water quality for agriculture, but would also improve the efficiency of water use in the catchment and enable progress to be made on water allocation reform as well as possibly even introduce avenues towards more equitable land redistribution. In this way, we have demonstrated how the WEF nexus approach can provide insights into how integrated water management can be applied in a particular agricultural context.

As water is a limiting factor in agriculture, especially given climate change and drought conditions which are anticipated to increase water stress, particularly on the agriculture sector, it is expedient to manage electricity costs through gravity-fed piped irrigation water, where possible, and to improve water quality in this way. It is thought that the option of gravity-fed piped water should be validated on a case-by-case basis, given the advantages of improved water quality, as well as the modelled advantages of enhanced profitability with reduced electricity usage, within this study. There are also many other, non-technical aspects that also need to be addressed on a case-specific basis, such as land ownership and access for pipeline routes, as well as environmental impact assessments.

Table 6: Effect on internal rate of return on capital investment (IRR, %) and net present value (NPV, ZAR) for selected water-related strategies for the 'typical farm' of each area on the Breede River, based on hypothetical scenarios

		Breedekloof		Worcester		Robertson		Montagu	
		IRR	NPV	IRR	NPV	IRR	NPV	IRR	NPV
Scenario 1 Land increases linked to water availability	-4 ha	2.79%	1 597 191	2.69%	995 126	3.96%	7 564 714	3.39%	2 744 678
	-2 ha	3.05%	2 525 493	3.02%	2 152 583	4.19%	8 679 302	3.76%	3 876 363
	Status quo	3.25%	3 453 796	3.35%	3 310 039	4.43%	9 793 890	4.12%	5 008 049
	+2 ha	3.48%	4 382 098	3.68%	4 467 495	4.67%	10 908 478	4.45%	6 139 734
	+4 ha	3.72%	5 310 401	4.01%	5 624 952	4.92%	12 023 066	4.77%	7 271 419
Scenario 2 Piped irrigation, with reductions in electricity	-30%	3.74%	5 434 118	3.76%	4 713 806	4.80%	11 561 291	4.63%	6 470 841
	-60%	4.40%	8 074 547	4.30%	6 585 494	5.31%	13 917 826	5.31%	8 421 232
	-100%	4.9%	10 054 868	4.77%	8 147 474	5.72%	15 869 271	5.83%	9 884 025
Scenario 3 Increase in water for white grape production	5% yield increase in white grapes	3.77%	5 525 972	3.78%	4 779 799	4.83%	11 682 554	5.31%	8 399 646

The insights into the WEF nexus were achieved by extending the focus on water quantity and quality to the high cost of using electricity to pump water, through a non-optimised, budget-based model. By addressing the energy cost, water use efficiency and water quality risks through an integrated dam and irrigation pipeline system, farm profitability could significantly increase, allowing farmers to look at new ways of addressing equity and land redistribution in the area. This gravity-fed closed system had several other projected benefits. It could reduce salinity in the system by limiting return flows off farms. It could free up additional fresh water from the Brandvlei Dam, positioned higher up in the catchment, that would have been used to flush out the additional salinity. Lastly, the piped system is an alternative to spending capital on reducing the leakages that were occurring in the current canal system, with greater benefits, in terms of the management of water quality and quantity through reduced evaporation levels. All these possibilities could be opened up by simply adopting a WEF nexus approach, rather than a singular focus on water.

Acknowledgements

This study was funded by the Western Cape Department of Environmental Affairs and Development Planning. We thank colleagues from Viticulture at the Agricultural Research Council in Nietvoorbij, Stellenbosch, the Robertson Winery in Robertson, farmers in Cogmanskloof as well as representatives from irrigation boards in Robertson and Cogmanskloof for sharing their knowledge.

Authors' contributions

L.S.: conceptualisation; data collection; literature review; writing the initial draft and revisions; project management. W.P.d.C.: conceptualisation; methodology; technical review and input; data analysis; funding acquisition. W.H.: conceptualisation; methodology; data collection; sample analysis; data analysis; writing the initial draft and revisions. J.D.S.C.: data analysis; writing the initial draft and revisions. A.M.H.: methodology; data collection; writing revisions; project leadership. M.d.W.: literature review; writing the initial draft and revisions; technical review.

References

1. Dillon JL, Hardaker JB. Farm management research for small farmer development. Rome: Food and Agriculture Organization; 1984.
2. Dorward PT, Shepherd DD, Wolmer WL. Developing farm management type methods for participatory needs assessment. Agric Syst. 1997;55(2):239–256. https://doi.org/10.1016/S0308-521X(97)00009-7
3. Keating BA, McCown RL. Advances in farming systems analysis and intervention. Agric Syst. 2001;70(2–3):555–579. https://doi.org/10.1016/S0308-521X(01)00059-2
4. Morgan-Davies C, Wilson R, Waterhouse T. Impacts of farmers' management styles on income and labour under alternative extensive land use scenarios. Agric Syst. 2017;155:168–178. https://doi.org/10.1016/j.agsy.2017.04.011
5. Leck H, Conway D, Bradshaw M, Rees J. Tracing the water–energy–food nexus: Description, theory and practice. Geogr Compass. 2015;9(8):445–460. https://doi.org/10.1111/gec3.12222
6. Wichelns D. The water-energy-food nexus: Is the increasing attention warranted, from either a research or policy perspective? Environ Sci Policy. 2017;69:113–123. https://doi.org/10.1016/j.envsci.2016.12.018
7. Kaddoura S, El Khatib S. Review of water-energy-food nexus tools to improve the nexus modelling approach for integrated policy making. Environ Sci Policy. 2017;77:114–121. https://doi.org/10.1016/j.envsci.2017.07.007
8. Den Hartigh W. Agriculture is still creating employment in SA. Farmers Weekly. 2016 May 11. Available from: https://www.farmersweekly.co.za/agri-news/south-africa/agriculture-is-still-creating-employment-in-sa/
9. Braham V, Dotwana A. Final land chapter. State of environment outlook report for the Western Cape Province 2014–2017 [document on the Internet]. c2018 [cited 2018 Aug 21]. Available from: https://www.westerncape.gov.za/eadp/files/atoms/files/03_Land.pdf
10. South African National Planning Commission. National Development Plan 2030. Our future – make it work. Pretoria: Government Printing Works; 2011.
11. Chambers D. Western Cape agriculture adds up the cost of the continuing drought. Business Day. 2018 July 19. Available from: https://www.businesslive.co.za/bd/national/2018-07-19-western-cape-agriculture-adds-up-the-cost-of-the-continuing-drought/
12. Green Cape. Market intelligence report [webpage on the Internet]. c2016 [cited 2018 Oct 08]. Available from: https://greencape.co.za/assets/GreenCape-Agriculture-MIR-2016
13. Oxford Dictionary [online]. Nexus [cited 2018 Mar 13]. Available from: https://en.oxforddictionaries.com/definition/nexus
14. Biggs EM, Boruff B, Bruce E, Duncan JMA, Haworth BJ, Duce S, et al. Environmental livelihood security in Southeast Asia and Oceania: A water-energy-food-livelihoods nexus approach for spatially assessing change. Colombo, Sri Lanka: International Water Management Institute; 2014. https://doi.org/10.5337/2014.231
15. Mehta L, Movik S. Flows and practices: Integrated water resources management (IWRM) in African contexts. IDS Working Papers. 2014;2014(438):1–34. https://doi.org/10.1111/j.2040-0209.2014.00438.x
16. Giordano M, Shah T. From IWRM back to integrated water resources management. Int J Water Resour Dev. 2014;30(3):364–376. https://doi.org/10.1080/07900627.2013.851521
17. Global Water Partnership. The need for an integrated approach [webpage on the Internet]. No date [cited 2018 Mar 13]. Available from: https://www.gwp.org/en/About/why/the-need-for-an-integrated-approach/
18. Claassen M. Integrated water resource management in South Africa. Int J Water Gov. 2013;1(3–4):323–338. https://doi.org/10.7564/13-IJWG12
19. Benson D, Gain A, Rouillard J. Water governance in a comparative perspective: From IWRM to a 'nexus' approach? Water Altern. 2015;8(1):756–773.
20. De Loë RC, Patterson JJ. Rethinking water governance: Moving beyond water-centric perspectives in a connected and changing world. Nat Resour J. 2017;57:75.
21. Water Research Commission. Water energy food security [webpage on the Internet]. No date [cited 2018 Mar 13]. Available from: http://www.wrc.org.za/Pages/LH3-WaterEnergyFoodSecurity.aspx#
22. Von Bormann T, Gulati M. The food energy water nexus: Understanding South Africa's most urgent sustainability challenge. Johannesburg: WWF-SA; 2014.
23. Statistics South Africa. Statistics by place [webpage on the Internet]. No date [cited 2018 Mar 13]. Available from: http://www.statssa.gov.za/?page_id=964
24. South African Department of Water Affairs and Forestry (DWAF). Strategic framework for national water resource quality monitoring programmes. Report no. N/0000/REQ0204. Pretoria: Resource Quality Services, DWAF; 2004.
25. Maree D. Why Western Cape agriculture matters to the South African economy. Business Report. 2017 June 28. Available from: https://www.iol.co.za/business-report/why-western-cape-agriculture-matters-to-sa-economy-10007415
26. Visser M, Ferrer S. Farm workers' living and working conditions in South Africa: Key trends, emergent issues, and underlying structural problems. Pretoria: International Labour Organization; 2015.
27. Erasmus D. More than just low wages behind De Doorns strike. Farmers Weekly. 2012 November 19. Available from: https://www.farmersweekly.co.za/agri-news/south-africa/more-than-just-low-wages-behind-de-doorns-strike/
28. Green Cape. Market intelligence report [webpage on the Internet]. c2017 [cited 2018 Mar 13]. Available from: https://www.greencape.co.za/assets/Uploads/GreenCape-Agri-MIR-2017-electronic-FINAL-v1.pdf
29. Du Preez P. Impact of downgrade on agriculture [webpage on the Internet]. c2017 [cited 2018 Mar 13]. Available from: https://www.cover.co.za/impact-downgrade-agriculture/
30. Venter M, Grové B, Van der Stoep I. The optimisation of electricity and water use for sustainable management of irrigation farming systems. Water Research Commission report no. TT717/17. Pretoria: Water Research Commission; 2017. Available from: http://www.wrc.org.za/Knowledge%20Hub%20Documents/Research%20Reports/TT%20717_final%20web.pdf

31. Slabbert A. Eskom increase unaffordable – AgriSA. Grid defection will accelerate utility's death spiral. MoneyWeb. 2017 October 06. Available from: https://www.moneyweb.co.za/news/south-africa/eskom-increase-unaffordable-agrisa/

32. South African Department of Water Affairs and Forestry (DWAF). State of the rivers report: Rivers of the Breede Water Management Area. Pretoria: DWAF; 2011. Available from: http://www.dwaf.gov.za/iwqs/rhp/state_of_rivers/WCape/Breede2011.pdf

33. Cullis JD, Gorgens AH, Marais C. A strategic study of the impact of invasive alien plants in the high rainfall catchments and riparian zones of South Africa on total surface water yield. Water SA. 2007;33(1):35–42. http://hdl.handle.net/10520/EJC116404

34. Jakku E, Thorburn PJ. A conceptual framework for guiding the participatory development of agricultural decision support systems. Agric Syst. 2010;103(9):675–682. https://doi.org/10.1016/j.agsy.2010.08.007

35. Cullis JDS, Rossouw N, Du Toit G, Petrie D, Wolfaardt G, De Clercq W, et al. Economic risks due to declining water quality in the Breede River Catchment. Water SA. 2018;44(3):464–473.

36. Kuehne G, Llewellyn R, Pannell DJ, Wilkinson R, Dolling P, Ouzman J, et al. Predicting farmer uptake of new agricultural practices: A tool for research, extension and policy. Agric Syst. 2017;156:115–125. https://doi.org/10.1016/j.agsy.2017.06.007

37. Daellenbach H, McNickle D, Dye S. Management science: Decision-making through systems thinking. Basingstoke: Palgrave Macmillan; 2012.

38. Nance RE, Sargent RG. Perspectives on the evolution of simulation. Oper Res. 2002;50(1):161–172. https://doi.org/10.1287/opre.50.1.161.17790

39. Strauss PG. Decision-making in agriculture: A farm-level modelling approach [MSc Agric thesis]. Pretoria: University of Pretoria; 2005.

40. Adamson D, Loch A, Schwabe K. Adaptation responses to increasing drought frequency. Aust J Agric Resour Econ. 2017;61(3):385–403. https://doi.org/10.1111/1467-8489.12214

41. Diogo V, Reidsma P, Schaap B, Andree BP, Koomen E. Assessing local and regional economic impacts of climatic extremes and feasibility of adaptation measures in Dutch arable farming systems. Agric Syst. 2017;157:216–229. https://doi.org/10.1016/j.agsy.2017.06.013

42. Jones JW, Antle JM, Basso B, Boote KJ, Conant RT, Foster I, et al. Brief history of agricultural systems modeling. Agric Syst. 2017;155:240–254. https://doi.org/10.1016/j.agsy.2016.05.014

43. Brenner T, Werker C. A taxonomy of inference in simulation models. Comput Econ. 2007;30(3):227–244. https://doi.org/10.1007/s10614-007-9102-6

44. Brennan T. Review of reasoning and method in economics: An introduction to economic methodology. J Econ Issues. 1981;15(3):796–799.

45. White AR. Inference. Philos Quart. 1971;21(85):289–302. https://doi.org/10.2307/2218655

46. Attonaty JM, Chatelin MH, Garcia F. Interactive simulation modeling in farm decision-making. Comput Electron Agr. 1999;22(2–3):157–170. https://doi.org/10.1016/S0168-1699(99)00015-0

47. Douthwaite B, Hoffecker E. Towards a complexity-aware theory of change for participatory research programs working within agricultural innovation systems. Agric Syst. 2017;155:88–102. https://doi.org/10.1016/j.agsy.2017.04.002

48. Robson AJ. The spreadsheet: How it has developed into a sophisticated modelling tool. Logist Inform Manage. 1994;7(1):17–23. https://doi.org/10.1108/09576059410052340

49. Rehman T, Dorward A. Farm management techniques and their relevance to administration, research and extension in agricultural development: Part 1 – Their evolution and use in developed countries. Agric Admin. 1984;15(3):177–189. https://doi.org/10.1016/0309-586X(84)90065-7

50. Malcolm LR. Fifty years of farm management in Australia: Survey and review. Rev Mark Agric Econ. 1990;58(1):24–55.

51. Pannell DJ. Lessons from a decade of whole-farm modeling in Western Australia. Rev Agric Econ. 1996;18(3):373–383. https://doi.org/10.2307/1349622

52. Doyle CJ. Application of systems theory to farm planning and control: Modelling resource allocation. Systems theory applied to agriculture and the food chain. London: Elsevier Applied Science; 1990. p. 89–112.

53. Blackie MJ, Dent JB. The concept and application of skeleton models in farm business analysis and planning. J Agric Econ. 1974;25(2):165–175. https://doi.org/10.1111/j.1477-9552.1974.tb00538.x

54. Carter HO. Representative farms: Guides for decision making? J Farm Econ. 1963;45(5):1448–1455. https://doi.org/10.2307/1236842

55. Feuz DM, Skold MD. Typical farm theory in agricultural research. J Sust Agr. 1992;2(2):43–58. https://doi.org/10.1300/J064v02n02_05

56. Ash A, Gleeson T, Hall M, Higgins A, Hopwood G, MacLeod N, et al. Irrigated agricultural development in northern Australia: Value-chain challenges and opportunities. Agric Syst. 2017;155:116–125. https://doi.org/10.1016/j.agsy.2017.04.010

57. Hortgro. Key deciduous fruit statistics. Paarl: Hortgro; 2016.

58. South African Wine Industry Information & Systems (SAWIS). SA wine industry 2016 statistics nr 41. Paarl: SAWIS; 2016.

59. VinPro. Cost guide 2017/2018. Paarl: VinPro Agricultural Economic Services; 2017.

60. South African Wine Industry Information & Systems (SAWIS). SA wine regions [document on the Internet]. c2017 [cited 2018 Aug 21]. Available from: http://www.sawis.co.za/cert/download/Regions_-_Aug_2017.pdf

61. CSIR. SA climate experts warn of a drier future in Western Cape [webpage on the Internet]. c2017 [cited 2018 Aug 21]. Available from: https://www.csir.co.za/sa-climate-experts-warn-drier-future-western-cape-0

62. Kienzle S, Flügel WA. The salinity of the Breede River and its tributaries between Brandvlei dam and H5M04: Summary of daily data up to September 1987. Pretoria: Hydrological Research Institute: Breede River Research Programme (BRSRP), Department of Water Affairs; 1988.

63. South African Department of Water Affairs and Forestry (DWAF). Breede water management area: Water resources assessment. Pretoria: DWAF; 2002.

Remains of a barn owl (*Tyto alba*) from the Dinaledi Chamber, Rising Star Cave, South Africa

AUTHORS:
Ashley Kruger[1]
Shaw Badenhorst[1]

AFFILIATION:
[1]Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Ashley Kruger

EMAIL:
ashleykruger@gmail.com

DATES:
Received: 06 June 2018
Revised: 27 July 2018
Accepted: 13 Sep. 2018
Published: 27 Nov. 2018

KEYWORDS:
Homo naledi; Cradle of Humankind; faunal remains; Pleistocene; vertebrate taphonomy

HOW TO CITE:
Kruger A, Badenhorst S. Remains of a barn owl (*Tyto alba*) from the Dinaledi Chamber, Rising Star Cave, South Africa. S Afr J Sci. 2018;114(11/12), Art. #5152, 5 pages. https://doi.org/10.17159/sajs.2018/5152

ARTICLE INCLUDES:
× Supplementary material
× Data set

FUNDING:
DST-NRF Centre of Excellence in Palaeosciences; Lee R Berger Foundation for Exploration; Lyda Hill Foundation; National Geographic Society; National Research Foundation (South Africa); University of the Witwatersrand

Excavations during November 2013 in the Rising Star Cave, South Africa, yielded more than 1550 specimens of a new hominin, *Homo naledi*. Four bird bones were collected from the surface of the Dinaledi Chamber during the first phase of the initial excavations. Although mentioned in the initial geological and taphonomic reports, the bird remains have not been formally identified and described until now. Here we identify these remains as the extant barn owl (*Tyto alba*) which is today common in the region and which is considered to have been an important agent of accumulation of microfaunal remains at many local Plio-Pleistocene sites in the Cradle of Humankind. Based on the greatest length measurement and breadth of the proximal articulation of the tarsometatarsus specimen, it is suggested that a single (female) individual is represented, despite the small sample sizes available for comparison. Although it is unclear how the remains of this female owl came to be accumulated in the remote Dinaledi Chamber, we suggest several possible taphonomic scenarios and hypothesise that these remains are not directly associated with the *Homo naledi* remains.

Significance:

- Owl bones from the Dinaledi Chamber are the only other macro-vertebrate remains from this Chamber.
- The other remains discovered are that of more than 15 individuals of the enigmatic *Homo naledi*.
- The remains of the Dinaledi Chamber owl further our understanding of the contents of the important material contained within the Dinaledi system as they are the only more recent fossils to be recovered from this area of the Rising Star Cave system and are therefore important in and of themselves as an indicator that more proximal parts of the Rising Star Cave system have been suitable for use by barn owls at greater time depths than the present.

Introduction

The Rising Star site

The Rising Star Cave system is located in the Cradle of Humankind UNESCO World Heritage Site, 50 km west-northwest of Johannesburg, South Africa (Figure 1). It is known that amateur cavers had periodically been visiting the cave system for a number of years (see Dirks et al.[1]); however, it was not until September 2013 that this system was formally investigated and fossil hominin remains were discovered in a very remote chamber named the Dinaledi Chamber.[1-3] Several excavations in the Chamber and adjacent spaces have yielded 1681 fossil hominin remains attributed to the new species *Homo naledi*.[1,2,4] Important to this study, approximately 300 bone specimens were collected from the cave surface of the Dinaledi Chamber and a further 1250 numbered fossil specimens were recovered from a small excavation pit in the cave floor no larger than 1 m^2 and less than 300 mm deep. This assemblage is the largest single collection of fossil hominin material found on the African continent to date, and the Rising Star Cave system is the only current location of remains of the hominin taxon, *H. naledi*.[1,2,4,5]

Dated to between 236 kya and 335 kya[5], geologically the Dinaledi Chamber and its fossil contents present an anomalous depositional environment in comparison to the 'classic' sites of the Cradle of Humankind in Gauteng

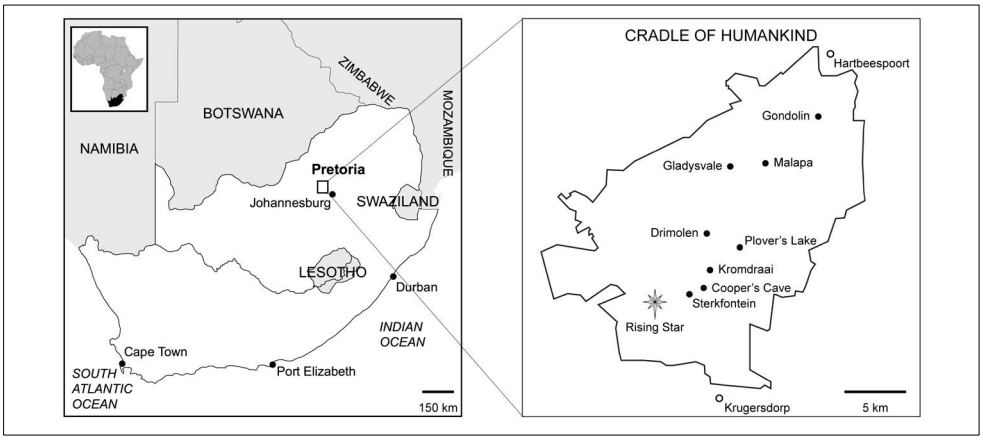


Figure 1: Location of the Rising Star Cave within the Cradle of Humankind UNESCO World Heritage Site. Other major palaeontological sites are also indicated.

Province, South Africa. Sites such as Sterkfontein, Kromdraai, Swartkrans and Malapa are noted for yielding fossil remains typically contained in lithified breccias, or found in decalcified sedimentary units derived ultimately from clastic lithified breccia.[6-11] In the majority of Cradle of Humankind fossil-bearing caves, it is hypothesised that skeletal material was brought into the system through a variety of agents, before being lithified.[12] Such agents can be biotic or abiotic, and include processes such as the effects of gravity (for example a fatal fall into a natural death trap, or downslope movement on talus slopes), vertebrate accumulation (predation or scavenging by carnivores, or accumulation by rodents), mass movement of sediments, fluvial transportation, or animal movement into the systems, or a combination of such processes. These processes generally produce taphonomic markers within a fossil assemblage, the role in site formation of which can be inferred from factors such as body-part representation, bone breakage patterns, or traces of surface modification including those of weathering, tooth marks or insect damage.[13-18]

Although a large number of fossil specimens of *H. naledi* have been recovered from the Dinaledi Chamber, surprisingly no definitive contemporaneous fauna[1] has been found to date, apart from four bird bones which were collected from the surface of the Dinaledi Chamber during excavations in 2013. It is clear from the first pictures taken by the exploration teams upon entering the Chamber in September of 2013 that these bones were placed together on a raised stone in the Chamber with a few other bones, indicating likely human agency (by explorers) in their positioning prior to discovery by scientists.

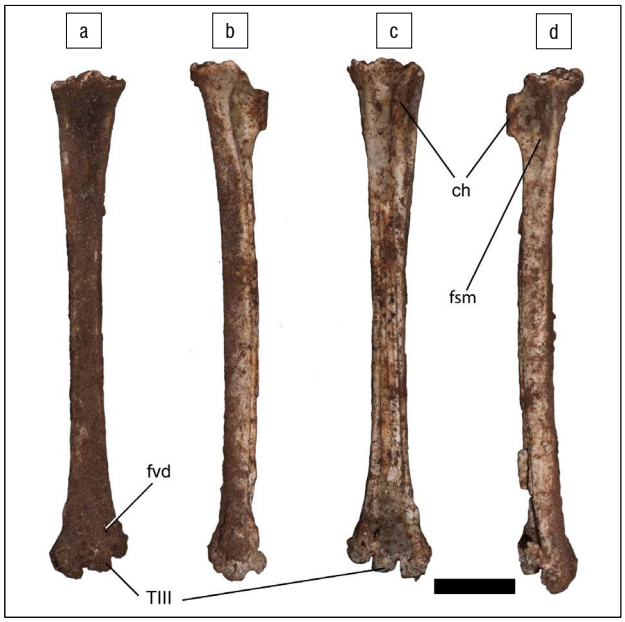
Based upon the physical state of the bones themselves as well as the clear lack of fossilisation as is typical on the hominin bones also found on the surface of the cave, we hypothesise that these remains are modern, or much closer to the present time, and not directly associated temporally with the *H. naledi* remains, likely being considerably younger. The appearance of preservation of these bones is clearly different from the hominin material found in the chamber.

While initially mentioned in the geological and taphonomic descriptions[1], the Dinaledi Chamber bird remains have not been formally identified or described until now. In this paper, we describe the four bird remains from the Dinaledi Chamber, using several possible explanations, expressed as hypotheses, to try explain how these bird bones were introduced into the Dinaledi Chamber.

Methods and results

The Dinaledi Chamber sample contains four bird specimens. Taxonomic diagnosis was made using comparative collections housed at the Ditsong National Museum of Natural History (formerly the Transvaal Museum) in Pretoria, South Africa. The measurements follow procedures given by Von den Driesch[19].

The most complete of these specimens, specimen U.W. 101 035 (Figure 2), a left tarsometatarsus, was used for skeletal measurements. This specimen is almost complete, with only the 3rd and 4th trochleas absent. The specimen is from an adult individual. The morphology of the proximal articulation distinctly places the specimen in the Strigiformes order, which consists of various species of owls. The morphology of the proximal articulation is identical to that of the extant barn owl (*Tyto alba*). This is supported by the greatest length of the specimen, which is also most similar to the barn owl (Table 1) among Strigiformes examined in this study. Very large- and small-sized owl species are not included because of the substantial adult size differences in these bones. In owls, the distal trochlea is similar in proportions. The presence of the 2nd trochlea is a reflection of maximum length.



Scale bar equals 1 cm

ch, crista medialis hypotarsi; fsm, facies subcutanea medialis; fvd, foramen vasculare distale; T, trochlea

Figure 2: Specimen U.W. 101 035, a left tarsometatarsus, is almost complete, with only the 3rd and 4th trochleas partially absent. The specimen is from an adult individual. (a) Dorsal, (b) lateral, (c) plantar and (d) medial views.

Table 1: Greatest length of tarsometatarsus (mm)

Taxon	Accession number	Greatest length of tarsometatarsus
Glaucidium perlatum (pearl spotted owl)	TM 71 880	22.74
Otus leucotis (white-faced owl)	TM 79 044	35.70
Strix woodfordii (wood owl)	TM 73 947	43.70
Asio capensis (march owl)	TM 80 555	51.84
Dinaledi specimen, identified as *Tyto alba*	U.W. 101 035	63.82
Bubo africanus (spotted eagle owl)	TM 76 097	69.62 (Note this species is much larger than the barn owl, as indicated by the measurements of this specimen: Bp: 12.52, SD: 6.46. See barn owl measurement in Table 2.)
Tyto capensis (grass owl)	TM 71 316	85.19

Bp, proximal breadth; SD, smallest breadth of the shaft

Based on measurement and morphology of the tarsometatarsus of the specimen from the Dinaledi Chamber, it is concluded that a barn owl (*Tyto alba*) is represented. In many bird species, male individuals are larger than female individuals, and this dimorphism can be reflected in anatomical measurements.[20] In the case of the barn owl, there is marked sexual dimorphism in terms of wing length (290–298 mm in male and 235–287 mm in female individuals[21]). Dimensions of the tarsometatarsus of specimen U.W. 101 035 from the Dinaledi Chamber appear to be more

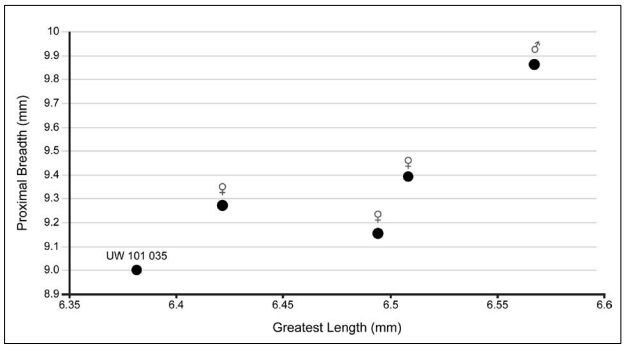


Figure 3: Tarsometatarsus proximal breath and greatest length measurements (in mm) for U.W. 101 035 plot closer to extant female specimens of *Tyto alba*.

Table 2: Barn owl tarsometatarsus measurements (mm) of male and female individuals

Specimen and sex	GL	Bp	SD	Bd
U.W. 101 035	63.82	9.0	4.01	10 (estimate)
TM 80 347 ♀	65.09	9.39	3.83	11.0
TM 80 242 ♀	64.22	9.27	4.08	11.36
TM 78 094 ♀	64.95	9.15	3.90	10.75
TM 80561 ♂	65.68	9.86	3.96	10.48

GL, greatest length; Bp, proximal breadth; SD, smallest breadth of the shaft; Bd, distal breadth

similar in size to that of female owls, but the comparative samples are small (Table 2; Figure 3).

A few other specimens accompanied the tarsometatarsus (Figure 4). These include: (1) a left tibiotarsus fragment of a bird, with the mid-distal shaft present, and part of the distal condyle (U.W. 101 40C; Figure 4a); (2) a right radius fragment of a bird, consisting of the distal, mid and proximal shaft, with the distal styloid process absent (U.W. 101 40B; Figure 4b); and (3) a right ulna fragment of a bird, consisting of the proximal, mid and distal shaft, with a small portion of the distal articulation present (U.W. 101 965 and 822; Figure 4c). In all cases, these specimens are similar in size and morphology to that of a barn owl, and are proportionally correct in size for an individual slightly smaller than the TM 80 242 and TM 78 094 specimens, and we conclude that they are most probably from the same individual as U.W. 101 035. However, the specimens are too fragmented to attempt to make an identification based on morphology alone. In this instance, the minimum number of individuals for *T. alba* is 1.

Discussion and conclusion

*The common barn owl (*Tyto alba*)*

Tyto alba is the most widely distributed species of owl in the world, and one of the most widespread of all birds, occupying many ecological areas, except Antarctica and parts of the Sahara Desert.[22] While almost exclusively nocturnal, in rare cases, *T. alba* is known to hunt diurnally.[23-25] As for most owls, the diet of *T. alba* comprises mainly small vertebrates, with a large majority represented by rodents. Undigested remains of the consumed prey in the stomachs of owls form into pellets, which are regurgitated, and are often rich in bone and teeth.[26-29] These regurgitated pellets are often found on the ground, underneath the diurnal resting area occupied by the owl.[27] A study done by Duke et al.[23] showed a striking difference in the bone composition of pellets when comparing those found in owls (48%) with those of hawks (6.5%) based on weight.

Scale bar equals 1 cm

Figure 4: (a) Specimen U.W. 101 40C, a left tibiotarsus fragment of *Tyto alba*, with the mid-distal shaft present, and part of the distal condyle; (b) specimen U.W. 101 40B, a right radius fragment of *Tyto alba*, consisting of the distal, mid and proximal shaft, with the distal styloid process absent; (c) specimens U.W. 101 965 and 822, a right ulna fragment of *Tyto alba*, consisting of the proximal, mid and distal shaft, with a small portion of the distal articulation present.

This richness and affinity for micromammal bone accumulation in the form of pellets are often found at palaeontological and archaeological sites around the world.

As the owl pellets fall to the ground, they slowly start disintegrating and start to be incorporated in the sediments. It is for this reason that owls are widely recognised as accumulating agents, and much work has been undertaken to examine the pellets of both modern and fossil owls[26,28,29], as the contents of these pellets serve as good palaeoenvironmental indicators[30-37]. In addition to serving as palaeoenvironmental indicators, owl pellets are also used in the study of stratigraphy at a particular site and the study of evolution of fauna.[27]

Barn owls are known to roost in a variety of habitats, occupying different types of cavity roosts. These habitats may include the twilight regions of rock fissures or hollow interiors of tree trunks (both dead and alive); owls are seldom found roosting in exposed roosts.[38] It is widely known that, in southern Africa, the protected openings and entrances of caves are often frequented by barn owls and used for roosting.[36,38-40]

Tyto alba is widespread in southern Africa and makes use of a variety of habitats ranging from woodlands to deserts, but excluding forests[21], and is known to roost in an assortment of places including cliffs, buildings, wells, mineshafts and caves. Owls have contributed microfauna remains to many Plio-Pleistocene sites in the Cradle of Humankind[41,42] including Gladysvale, Kromdraai, Sterkfontein and Swartkrans[29,37-39,43-48].

The Dinaledi Owl

Figure 5a shows the Dinaledi Chamber and the location, where recovered, of three of the four bird remains discussed here. As noted earlier, the remains had apparently been picked up and placed on a rock in the distal section of the Dinaledi Chamber (Figure 5b) by one or more cavers prior to the exploration of the Chamber by our scientific team. As noted by Dirks et al.[1], the owl remains are taphonomically distinct from the rest of the hominin assemblage as they lack surface modification and stain patterns seen on the hominin remains; in addition, they are covered in an adhesion of a thin film of calcite crystals. These crystals cover much of the surface of the owl bone, and thus suggest that they may have been deposited relatively recently.

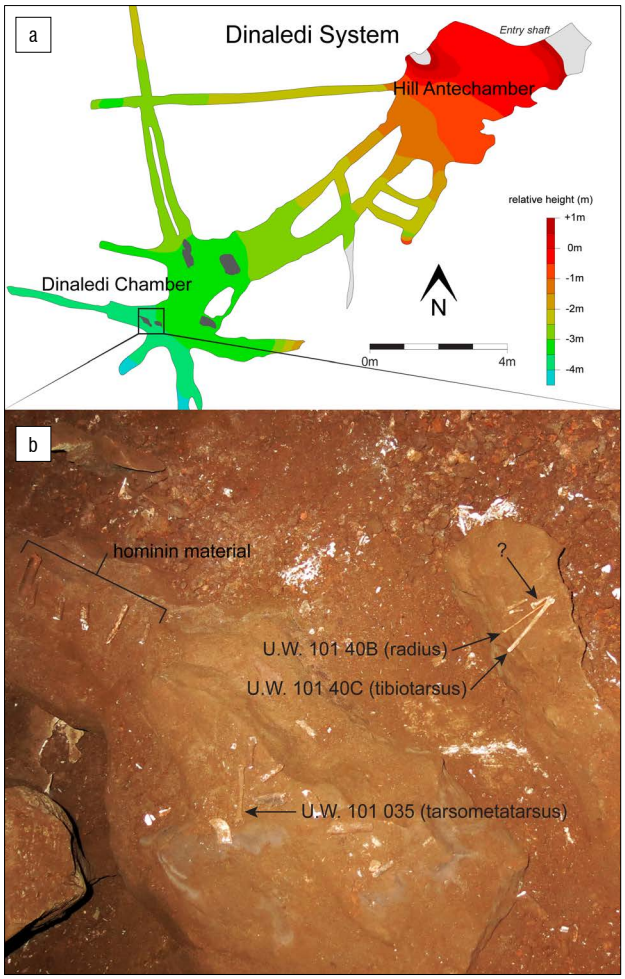


Figure 5: (a) Map of the Dinaledi Chamber in the Dinaledi system, Rising Star Cave (modified after Dirks et al.[1]); (b) location of three of the four owl remains found within the Chamber. The remains were seemingly placed on a rock, amongst other hominin material, by an unknown caver prior to investigations by the University of the Witwatersrand. An unidentifiable fragment (denoted by ?) is possibly remains of a bird, but lacks enough morphology for a proper diagnosis.

There are several possible scenarios to explain the placement of these bird remains. A modern owl may have become lost in the system and, by way of flying around in the dark, found its way into the Dinaledi Chamber (Hypothesis 1). It is also possible that the remains fell down the narrow chute – a 12-m drop immediately above the Dinaledi system, and the only currently known accessible route into this system (Hypothesis 2). If this were the case, it is possible that more remains from this individual are still to be recovered, possibly from the base of the chute, in what is now called the Hill Antechamber of the Dinaledi system[49], although in this scenario, human agency would be required for the bones to find their way into the more distal Dinaledi Chamber.

Alternatively, but less likely, the remains of this modern owl could have been introduced into the system by a caver carrying them in, possibly from the main entrance of the Rising Star Cave (Hypothesis 3). This hypothesis is supported by the fact that the four remains were found on a rock within the Dinaledi Chamber, although why a caver would carry remains into such a difficult to access area and then leave them there is not obvious. In this scenario, it is possible that the owl died at the entrance of the Rising Star Cave, as this area has a constructed entrance (created by limestone miners at some point in the 1930s) and a natural cave roof opening, both of which open up into a rocky cave wall approximately 6 m high. This rock face is a suitable roosting area, and is currently inhabited by bats, swallows and other bird species including occasionally barn owls. This open area has a relative abundance of light, and as is commonly known, barn owls will seek out roosts which are dark and enclosed, even when roosting in trees or on the ground.[38]

An alternative entrance into the Dinaledi Chamber, as suggested for example by Thackeray[50], is another possible scenario for the introduction of the modern owl remains (Hypothesis 4). However, extensive and exhaustive exploration by cavers from the University of the Witwatersrand has, to date, failed to identify another entrance to the Dinaledi system.

We favour one of the first two hypotheses as the most likely origin of this material in the position of their discovery: the owl became lost in the system and found its way either into the Dinaledi Chamber, or to the top of the chute, before perishing.

The owl remains from the Dinaledi Chamber help further our understanding of the contents of the important material contained within the Dinaledi system. They are also the only more recent fossils to be recovered from this area of the Rising Star Cave system and are therefore important in and of themselves as an indicator that more proximal parts of the Rising Star Cave system have been suitable for use by barn owls at greater time depths than the present. Once described, some of these bones will be subjected to radiocarbon dating to establish if they might be useful in placing an uppermost date on the bone content of the Dinaledi Chamber as perhaps they date the last depositional event to occur within the relatively closed Dinaledi system.

Acknowledgements

We acknowledge funding from: DST-NRF Centre of Excellence in Palaeosciences (CoE-Pal); Lee R Berger Foundation for Exploration; Lyda Hill Foundation; National Geographic Society; National Research Foundation (South Africa); and University of the Witwatersrand. The support of the CoE-Pal towards this research is hereby acknowledged. Opinions expressed, and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the CoE-Pal. We thank B. Zipfel and S. Jirah for access to collections at the Evolutionary Studies Institute. We thank the Ditsong National Museum of Natural History for access to their faunal collection.

Authors' contributions

A.K. conducted the research and wrote the original draft of the manuscript. S.B. took all measurements and undertook the identification. Both authors contributed equally to manuscript revisions and editing.

References

1. Dirks PHGM, Berger LR, Roberts EM, Kramers JD, Hawks J, Randolph-Quinney PS, et al. Geological and taphonomic context for the new hominin species *Homo naledi* from the Dinaledi Chamber, South Africa. eLife. 2015;4, e09561, 37 pages. https://doi.org/10.7554/eLife.09561
2. Berger LR, Hawks J, De Ruiter DJ, Churchill SE, Schmid P, Delezene LK, et al. *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. eLife. 2015;4, e09560, 35 pages. https://doi.org/10.7554/eLife.09560
3. Kruger A, Randolph-Quinney PS, Elliott M. Multimodal spatial mapping and visualisation of Dinaledi Chamber and Rising Star Cave. S Afr J Sci. 2016;112(5/6), Art. #2016-0032, 11 pages. http://dx.doi.org/10.17159/sajs.2016/20160032

4. Hawks J, Elliott M, Schmid P, Churchill SE, Ruiter DJd, Roberts EM, et al. New fossil remains of *Homo naledi* from the Lesedi Chamber, South Africa. eLife. 2017;6, e24232, 63 pages. https://doi.org/10.7554/eLife.24232

5. Dirks PHGM, Roberts EM, Hilbert-Wolf H, Kramers JD, Hawks J, Dosseto A, et al. The age of *Homo naledi* and associated sediments in the Rising Star Cave, South Africa. eLife. 2017;6, e24231, 59 pages. https://doi.org/10.7554/eLife.24231

6. Clarke RJ. On some new interpretations of Sterkfontein stratigraphy. S Afr J Sci. 1994;90:211–214.

7. Partridge TC. Re-appraisal of lithostratigraphy of Sterkfontein hominid site. Nature. 1978;275:282–287. https://doi.org/10.1038/275282a0

8. Berger LR, Menter CG, Thackeray JF. The renewal of excavation activities at Kromdraai, South Africa. S Afr J Sci. 1994;90:209–210.

9. Kuman K, Field AS, Thackeray JF. Discovery of new artefacts at Kromdraai. S Afr J Sci. 1997;93:187–193.

10. Brain CK. Structure and stratigraphy of the Swartkrans Cave in the light of the new excavations. In: Brain CK, editor. Monograph 8: Swartkrans. Pretoria: Transvaal Museum; 1993. p. 23–34.

11. De Ruiter DJ. Revised faunal lists for Members 1–3 of Swartkrans, South Africa. Ann Transv Mus. 2003;40:29–41.

12. Dirks PHGM, Berger LR. Hominin-bearing caves and landscape dynamics in the Cradle of Humankind, South Africa. J Afr Earth Sci. 2013;78:109–131. https://doi.org/10.1016/j.jafrearsci.2012.09.012

13. Lyman R. Vertebrate taphonomy. Cambridge: Cambridge University Press; 1994. https://doi.org/10.1017/CBO9781139878302

14. Lyman R, Fox G. A critical evaluation of bone weathering as an indication of bone assemblage formation. In: Haglund W, Sorg M, editors. Forensic taphonomy: The post-mortem fate of human remains. Boca Raton, FL: CRC Press; 1997. p. 223–247.

15. Lyman RL. What taphonomy is, what it isn't, and why taphonomists should care about the difference. J Taphonomy. 2010;8(1):1–16.

16. Pokines JT. Faunal dispersal, reconcentration, and gnawing damage to bone in terrestrial environments. In: Pokines J, Symes S, editors. Manual of forensic taphonomy. Boca Raton, FL: CRC Press; 2013. p. 201–248.

17. Pokines JT, Baker JE. Effects of burial environment on osseous remains. In: Pokines J, Symes S, editors. Manual of forensic taphonomy. Boca Raton, FL: CRC Press; 2013. p. 73–114. https://doi.org/10.1201/b15424-6

18. Bristow J, Simms Z, Randolph-Quinney PS. Taphonomy. In: Black S, Ferguson E, editors. Forensic anthropology 2000-2010. Boca Raton, FL: CRC Press; 2011. p. 279–318.

19. Von den Driesch A. A guide to the measurement of animal bones from archaeological sites: As developed by the Institut für Palaeoanatomie, Domestikationsforschung und Geschichte der Tiermedizin of the University of Munich. Cambridge, MA: Peabody Museum Press; 1976.

20. Badenhorst S, Lyle R, Merewether J, Driver J, Ryan S. The potential of osteometric data for comprehensive studies of turkey (*Meleagris gallopavo*) husbandry in the American Southwest. Kiva. 2012;78(1):61–78. https://doi.org/10.1179/kiv.2012.78.1.61

21. Maclean GL. Robert's birds of southern Africa. Cape Town: John Voelcker Bird Book Fund; 1985.

22. Shawyer CR. The barn owl. London: Hamlyn; 1994.

23. Duke G, Jegers A, Loff G, Evanson O. Gastric digestion in some raptors. Comp Biochem Physiol A Physiol. 1975;50(4):649–656. https://doi.org/10.1016/0300-9629(75)90121-8

24. Bunn D. Regular daylight hunting by barn owls. Brit Birds. 1972;65:26–30.

25. Harte K. Barn owl hunting by daylight. The Wilson Bulletin. 1954:270–270.

26. Dodson P, Wexlar D. Taphonomic investigations of owl pellets. Paleobiology. 1979;5(3):275–284. https://doi.org/10.1017/S0094837300006564

27. Culver DC, White WB. Encyclopedia of caves. Amsterdam: Elsevier; 2005.

28. Kusmer KD. Taphonomy of owl pellet deposition. J Paleontol. 2015;64(4):629–637. https://10.1017/S0022336000042669

29. Levinson M. Taphonomy of microvertebrates – from owl pellets to cave breccia. Ann Transv Mus. 1982;33(6):115–121.

30. Winkler AJ. Neogene paleobiogeography and East African paleoenvironments: Contributions from the Tugen Hills rodents and lagomorphs. J Hum Evol. 2002;42(1–2):237–256. https://doi.org/10.1006/jhev.2001.0501

31. Jaeger J, Wesselman H. Fossil remains of micromammals from the Omo Group deposits. Earliest man and environments in the Lake Rudolf Basin. Chicago, IL: University of Chicago Press; 1976. p. 351–360.

32. Fernandez-Jalvo Y, Andrews P. Small mammal taphonomy of Gran Dolina, Atapuerca (Burgos), Spain. J Archaeol Sci. 1992;19(4):407–428. https://doi.org/10.1016/0305-4403(92)90058-B

33. Denys C. Fossil rodents (other than Pedetidae) from Laetoli. In: Leakey MD, Harris JM, editors. Laetoli: A Pliocene site in northern Tanzania. Oxford: Oxford University Press; 1987. p. 118–170.

34. De Graaff G. On the fossil mammalian microfauna collected at Kromdraai by Draper in 1895. S Afr J Sci. 1961;57(9):259–260.

35. Avery D. The environment of early modern humans at Border Cave, South Africa: Micromammalian evidence. Palaeogeogr Palaeoclim Palaeoecol. 1992;91(1–2):71–87. https://doi.org/10.1016/0031-0182(92)90033-2

36. Davis D. The barn owl's contribution to ecology and palaeoecology. Ostrich. 1959;30(S1):144–153. https://doi.org/10.1080/00306525.1959.9633322

37. Avery D. The Plio-Pleistocene vegetation and climate of Sterkfontein and Swartkrans, South Africa, based on micromammals. J Hum Evol. 2001;41(2):113–132. https://doi.org/10.1006/jhev.2001.0483

38. Reed DN. Taphonomic implications of roosting behavior and trophic habits in two species of African owl. J Archaeol Sci. 2005;32(11):1669–1676. https://doi.org/10.1016/j.jas.2005.05.007

39. Reed DN. Micromammal paleoecology: Past and present relationships between African small mammals and their habitats. Stony Brook, NY: State University of New York; 2003.

40. McCrae C. A comparative study of Late Holocene and Plio-Pleistocene-aged micromammalian owl accumulations from South Africa. Palaeontol Afr. 2009;44:190–191.

41. Brain CK. The hunters or the hunted? Chicago, IL: Chicago University Press; 1981.

42. Watson V. Form, function and fibres: A preliminary study of the Swartkrans fossil birds. Koedoe. 1991;34(1):23–29. https://doi.org/10.4102/koedoe.v34i1.410

43. Glue DE. Avian predator pellet analysis and the mammalogist. Mammal Rev. 1970;1(3):53–62. https://doi.org/10.1111/j.1365-2907.1970.tb00320.x

44. Vernon C. An analysis of owl pellets collected in southern Africa. Ostrich. 1972;43(2):109–124. https://doi.org/10.1080/00306525.1972.9632586

45. Pocock T. Plio-Pleistocene fossil mammalian microfauna of southern Africa – A preliminary report including description of two new fossil muroid genera (Mammalia: Rodentia). Palaeontol Afr. 1987;26(7):69–71.

46. Avery D. A preliminary assessment of the micromammalian remains from Gladysvale Cave, South Africa. Palaeontol Afr. 1995;32:1–10.

47. Avery D. Micromammals as palaeoenvironmental indicators of the southern African Quaternary. T Roy Soc S Afr. 2007;62(1):17–23. https://doi.org/10.1080/00359190709519193

48. Leichliter JN. Micromammal paleoecology: Theory, methods, and application to modern and fossil assemblages in the Cradle of Humankind World Heritage Site, South Africa. Boulder, CO: University of Colorado Boulder; 2011.

49. Berger LR, Elliott MC, Peixotto B, Morris H, Feuerriegel EM, Tucker SJ, et al. A new naming scheme for the Dinaledi Chamber System and associated antechambers and passages of the Rising Star Cave System, South Africa. Paper presented at: The 87th Annual Meeting of the American Association of Physical Anthropologists; 2018 April 11–14; Austin, Texas, USA. Am J Phys Anthropol. 2018;165(S66):25–26.

50. Thackeray JF. The possibility of lichen growth on bones of *Homo naledi*: Were they exposed to light? S Afr J Sci. 2016;112(7–8), Art. #a0167, 5 pages. http://dx.doi.org/10.17159/sajs.2016/a0167

Hominin cranial fragments from Milner Hall, Sterkfontein, South Africa

AUTHORS:
Amélie Beaudet[1,2]
Jason L. Heaton[3,4,5]
Ericka N. L'Abbé[2]
Travis R. Pickering[4,5,6]
Dominic Stratford[1]

AFFILIATIONS:
[1]School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa
[2]Department of Anatomy, University of Pretoria, Pretoria, South Africa
[3]Department of Biology, Birmingham-Southern College, Birmingham, Alabama, USA
[4]Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa
[5]Plio-Pleistocene Palaeontology Section, Department of Vertebrates, Ditsong National Museum of Natural History, Pretoria, South Africa
[6]Department of Anthropology, University of Wisconsin, Madison, Wisconsin, USA

CORRESPONDENCE TO:
Amélie Beaudet

EMAIL:
beaudet.amelie@gmail.com

DATES:
Received: 01 July 2018
Revised: 15 Aug. 2018
Accepted: 15 Aug. 2018
Published: 27 Nov. 2018

KEYWORDS:
Cradle of Humankind; late Pliocene–early Pleistocene; cranial thickness; diploë; *Homo*

HOW TO CITE:
Beaudet A, Heaton JL, L'Abbé EN, Pickering TR, Stratford D. Hominin cranial fragments from Milner Hall, Sterkfontein, South Africa. S Afr J Sci. 2018;114(11/12), Art. #5262, 6 pages. https://doi.org/10.17159/sajs.2018/5262

ARTICLE INCLUDES:
× Supplementary material
× Data set

The Sterkfontein Caves site is one of the richest early hominin localities in Africa. In addition to significant fossil assemblages from Members 2 and 4 of the Sterkfontein Formation, recent excavations have revealed hominin-bearing sedimentary deposits in the lesser-known Milner Hall. We describe two hominin cranial fragments excavated from the Milner Hall in 2015 and present the results of a high-resolution microtomographic-based approach to diagnosing the anatomical and taxonomical origin of these specimens. Based on external morphology, StW 671 and StW 672 are identified as frontal and occipital fragments, respectively. Our non-invasive bi-dimensional quantitative investigation of the two cranial fragments reveals a mean cranial thickness of 8.8 mm for StW 671 and of 5.6 mm for StW 672, and a contribution of the diploic layer to the cumulative cranial thickness that is less than 50%. While the mean cranial thickness of StW 671 falls within the range for *Homo*, the relative proportion of the diploë in both StW 671 and StW 672 is lower than that found in *Australopithecus* (>60%) and extant humans (>50%). Accordingly, in terms of both cranial thickness and inner structural organisation, the Milner Hall hominins combine derived and unique traits, consistent with the condition of other postcranial and dental material already described from the deposit. Moreover, our study opens interesting perspectives in terms of analysis of isolated cranial fragments, which are abundant in the hominin fossil record.

Significance:

- The Sterkfontein Caves have widely contributed to our understanding of human evolution.
- Besides the well-known Members 4 and 2, where the iconic 'Mrs Ples' and 'Little Foot' have been found, in this study we suggest that the Milner Hall locality represents an additional, stratigraphically associated source of not only fossil hominins, but also Oldowan stone tools.
- In particular, we describe for the first time two cranial fragments, StW 671 and StW 672, identified as frontal and occipital bones, respectively.
- Our microtomographic-based analysis of these materials reveals some affinities with *Homo* combined with unique characters.
- In this context, our study suggests an intriguing mosaicism consistent with the description of the two fossil hominins found in the Milner Hall.

Introduction

Milner Hall (MH) is a deep underground chamber of the Sterkfontein Caves (South Africa) that extends about 100 m in a roughly east-west direction. Two hominin fossils, a molar and a proximal phalanx, excavated from the T1 depositional unit of the Central Underground Deposits excavation site in the Milner Hall (STK-MH1) have already been described.[1] The complex stratigraphic context of the Milner Hall fossiliferous depositional sequence, in which an early distal accumulation of the 3.67-Ma-old Member 2 (T3) and 2.18-Ma-old Oldowan artefact-bearing sediments from Member 5 (T2) contribute to the formation of T1[2,3], affords that potentially *Australopithecus*, *Paranthropus* and early *Homo* might be represented in the unit. Interestingly, the description and metric analyses of the two first hominin specimens excavated from T1 suggest an enigmatic mix of unique, primitive and derived morphological traits, with potential morphological affinities with the genus *Homo*.[1]

With hominin remains dated potentially to either 3.67 Ma or 2.18 Ma, the Milner Hall fossil assemblage may contribute to ongoing debates about hominin morphological and taxonomic diversity at Sterkfontein during the late Pliocene and early Pleistocene. In particular, because of the poorly known degree of intraspecific variation in the *Australopithecus* hypodigm, the Sterkfontein hominin fossil assemblages have been the focus of long-standing discussions regarding the presence of one or two *Australopithecus* species.[4,5] In addition, the species-level diagnosis of purported early *Homo* remains from Member 5 of the Sterkfontein Formation is contentious.[6-8] In particular, the re-attribution of hominin remains previously assigned to early *Homo* or *Australopithecus* (e.g. StW 53)[8], and the fragmentary nature of fossil specimens identified as early *Homo* (e.g. SK 847)[9], complicate our understanding of early human diversity in South Africa. In this context, the Milner Hall deposits have the potential to provide further evidence to critically assess hominin palaeobiodiversity and the taxonomic context of the Sterkfontein hominin-bearing deposits.

FUNDING:
AESOP+ Programme; Claude Leon Foundation; DST-NRF Centre of Excellence in Palaeosciences; French Institute of South Africa; NRF African Origins Platform; Palaeontological Scientific Trust (PAST)

Here, we report on two additional hominin fossils excavated from STK-MH1 and discuss efforts to assign them taxonomically.

Material and methods

As comparative material, we investigated four South African hominin cranial specimens from Sterkfontein (Sts 5, Sts 71, StW 505) and Swartkrans (SK 46), attributed, respectively, to *Australopithecus africanus* and *Paranthropus robustus*.[10-14] In the absence of well-preserved South African early Pleistocene human crania, we included the Middle Pleistocene composite cranium DH 1/DH 3 from Rising Star attributed to *Homo naledi* as comparative material for the external morphology.[15] Additionally, we selected bone thickness values of nine human and non-human Pliocene and Pleistocene hominin taxa presented in the supplementary information of Copes and Kimbel[16]. Our extant comparative sample comprises adult humans (*Homo sapiens*, $n=10$) and common chimpanzees (*Pan troglodytes*, $n=10$) with equal proportions of male and female individuals within each taxon from the collections of the University of Pretoria[17] (South Africa) and the Royal Museum for Central Africa (Belgium) respectively. Ethical clearance for the use of extant human crania was obtained from the Main Research Ethics Committee of the Faculty of Health Sciences, University of Pretoria in February 2016 (35/2016).

The new fossils, StW 671 and StW 672, were scanned at the microfocus X-ray tomography facility of the Palaeosciences Centre at the University of the Witwatersrand, in Johannesburg (South Africa), at a spatial resolution of 28 μm (isotropic voxel size) (Figures 1 and 2). Fossil comparative material from Sterkfontein and Swartkrans as well as extant specimens were scanned at the Palaeosciences Centre, at the South African Nuclear Energy Corporation in Pelindaba (South Africa), or at the Centre for X-ray Tomography of the Ghent University in Ghent (Belgium). Additionally, for external morphology, we included in our study the digital replica of the composite cranium DH 1/DH 3 from Rising Star available on MorphoSource (www.morphosource.org).[15]

StW 671 and StW 672 were digitally rendered with Avizo v.9.0 software (Visualization Sciences Group Inc.). We extracted one section in StW 671 and one section in StW 672, sampling maximum length/width of the preserved cranial fragments (Figure 2). In terms of measurements, we followed the protocol detailed in Beaudet et al.[18] In view of collecting consistent data throughout the sample, we tried to avoid oversampling exocranial reliefs and, thus, the section in StW 671 was positioned orthogonally to the temporal line. The three layers of the bone were segmented by combining manual and automatic methods (i.e. watershed transform). Thickness and surface area of the inner table of the diploë and of the outer table were automatically and separately measured at regular intervals along the two sections perpendicular to the outer cranial surface using a custom-written program in MATLAB R2013a[18] (Mathworks, https://www.mathworks.com/products/matlab.html) (Figure 2). In total, 30 measurements were collected on StW 671 and on StW 672. Based on surface area, we computed bone tissue proportions as the percentage of the bone area represented by outer/inner tables or diploë.[18] For comparative specimens, with the exception of DH 1/DH 3, we selected a number of measurements (see Results) performed along parasagittal sections from Beaudet et al.[18] and corresponding to the portions of the cranium documented in StW 671 and StW 672.

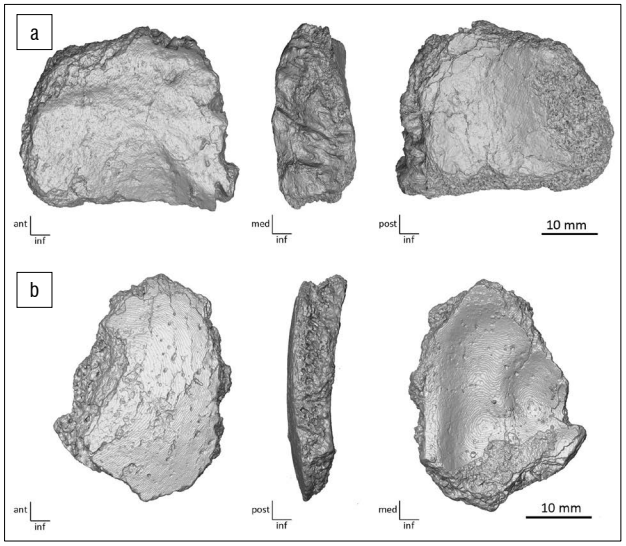


Figure 1: Virtual rendering of (a) StW 671 and (b) StW 672 in ectocranial (left), posterior/medial (middle) and endocranial (right) views.

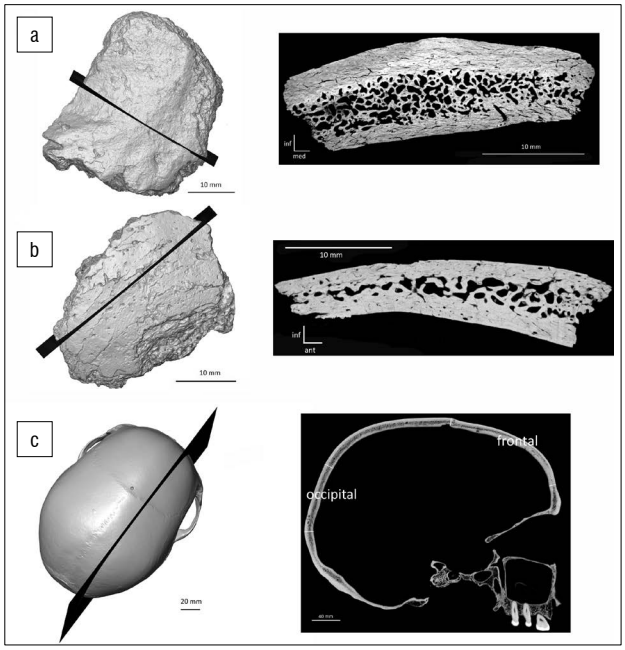


Figure 2: Cranial structural organisation along sections extracted from (a) StW 671 and (b) StW 672. Comparative measurements come from (c) the frontal and occipital parts of the parasagittal section extracted from the comparative sample. White lines in (c) delimitate the frontal and occipital portions of the cranial vault.

Results

Description of StW 671 and StW 672

StW 671 is a small rectangular fragment of the frontal bone (maximum length = 38 mm; maximum width = 32.5 mm; maximum intertabular thickness = 13.7 mm) (Figure 1a). On the lateral aspect, the temporal line runs along the ectocranial surface. The posterior border of the fragment corresponds to the coronal suture. The surface immediately inferior to the temporal line is relatively flat. The endocranial surface is convex and does not exhibit any diagnostic anatomical features. In terms of surface preservation, StW 671 shows a few minor weathering cracks and heavy manganese dioxide staining.

StW 672 is an ovoid fragment of the occipital bone (maximum length = 34.3 mm; maximum width = 25.1 mm; maximum thickness = 7.4 mm) (Figure 1b). The ectocranial surface is relatively smooth. The endocranial surface preserves two shallow grooves and a faint ridge is perceptible along the medial border. The fragment is globally convex. Because the ectocranial and endocranial surfaces lack crests and vascular grooves, respectively, this fragment is proposed to originate from the portion of the occipital bone superior to the nuchal crest and transverse sinus. The specimen's surface shows minor weathering, including cracks, and evidence of diagenetic flaking in the form of non-morphological pits on the ectocranial aspect.

Comparative description of StW 671 and StW 672

StW 671 and StW 672 were superimposed on the digital replica of Sts 5 (*Australopithecus africanus*), SK 46 (*Paranthropus robustus*), a composite skull of *Homo naledi* based on DH 1 and DH 3, and extant human and chimpanzee crania (Figure 3).

The overall morphology of StW 671 does not fit well with the external aspect of the frontal bone of Sts 5, SK 46, DH 1/DH 3 and of the chimpanzee specimen. Indeed, when superimposed onto Sts 5 and DH 1/DH 3, the portion of the frontal bone above the temporal line in StW 671 is more elevated than in the comparative frontal squama. The temporal line progressively joins the sagittal crest in *Paranthropus* and chimpanzees.[19] If the temporal line of StW 671 is superimposed onto the temporal line of SK 46 and of the chimpanzee specimen, the coronal suture in the Milner Hall specimen is then positioned in the middle of the frontal bone of the two comparative specimens. The overall morphology of the StW 671 fragment closely fits the shape of the frontal bone in the extant human specimen. However, when superimposed, the temporal line in StW 671 runs inferiorly compared to the human cranium.

The morphology of StW 672 is compatible with the morphology of the occipital bone for all comparative crania considered in this study.

Cranial thickness and composition in StW 671 and StW 672

Thickness values and tabular proportions in StW 671 and StW 672 as well as in the comparative sample are shown in Tables 1 and 2 and in Figure 4. From Beaudet et al.[18], we specifically selected measurements 1–20 and 40–50, respectively representing the frontal and occipital bones (Figure 2c). Additionally, we compiled comparative measurements of the frontal bone of non-human and human hominin taxa from Copes and Kimbel[16] in Table 3.

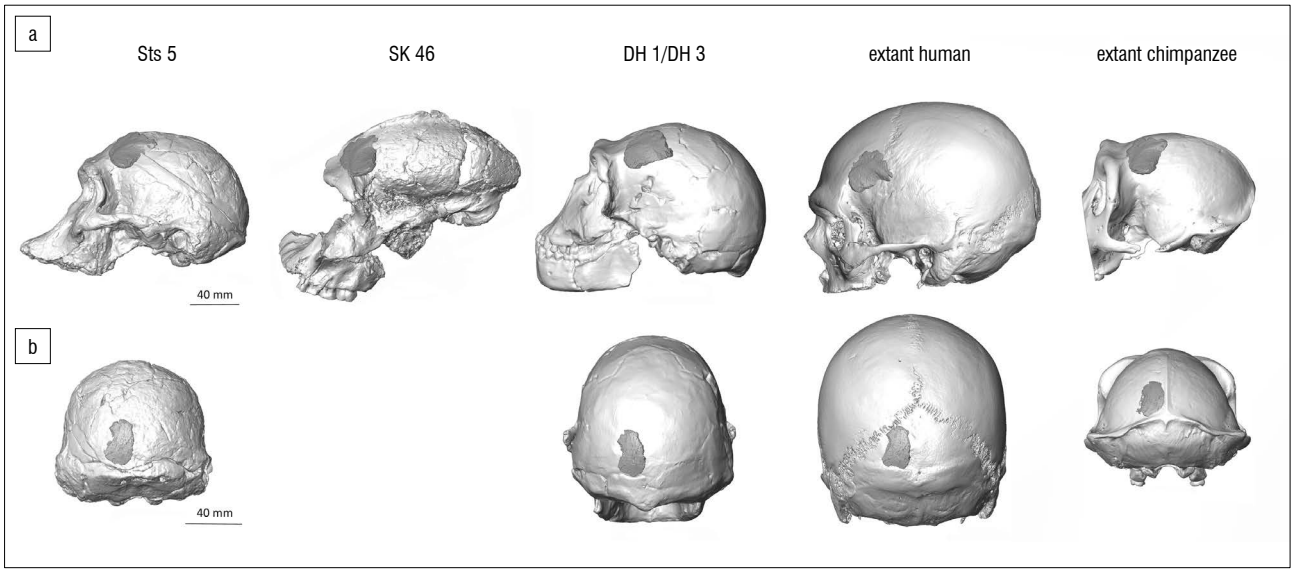


Figure 3: Superimposition of (a) StW 671 (in lateral view) and (b) StW 672 (in posterior view) on the digital replicas of Sts 5 (*Australopithecus africanus*), SK 46 (*Paranthropus robustus*), DH 1/DH 3 (*Homo naledi*) and extant human and chimpanzee crania.

Table 1: Mean frontal bone thickness (FBT, mm) and relative proportions of the diploë (%) of StW 671 compared to the estimates from some Plio-Pleistocene hominin specimens and extant humans and chimpanzees. Standard deviations are shown in brackets.

	FBT	Outer table	Inner table	Diploë	% Diploë
StW 671	8.8 (0.9)	2.9 (0.3)	1.8 (0.4)	4.1 (0.8)	45.1 (5.9)
StW 505	10.2 (5.3)	–	–	–	–
Sts 5	6.3 (3.1)	0.2 (0.3)	1.1 (0.6)	5.0 (2.5)	77.6 (4.1)
Sts 71	7.8 (1.1)	0.9 (0.2)	0.6 (0.3)	6.3 (1.2)	78.7 (5.3)
SK 46	7.4 (4.2)	7.1 (4.2)*		0.3 (0.5)	6.4 (8.9)
Extant humans (n=10)	7.6 (0.7)	1.9 (0.4)	1.7 (0.2)	4.1 (0.8)	51.0 (5.9)
Extant chimpanzees (n=10)	4.6 (1.0)	3.7 (0.5)*		1.0 (0.8)	15.6 (4.7)

**The outer and inner tables are indistinct (i.e. cortical) in SK 46 and extant chimpanzees.*

Table 2: Mean occipital bone thickness (OBT, mm) and relative proportions of the diploë (%) of StW 672 compared to the estimates from some Plio-Pleistocene hominin specimens and extant humans and chimpanzees. Standard deviations are shown in brackets.

	OBT	Outer table	Inner table	Diploë	% Diploë
StW 672	5.6 (0.9)	1.4 (0.3)	1.8 (0.3)	2.4 (0.9)	41.3 (10.2)
StW 505	9.3 (4.9)	–	–	–	–
Sts 5	9.2 (4.2)	0.6 (0.6)	0.9 (0.3)	7.7 (3.9)	83.4 (35.6)
Sts 71	7.0 (2.9)	0.2 (0.3)	0.3 (0.2)	6.4 (2.8)	89.5 (43.6)
SK 46	6.7 (1.3)	6.1 (1.8)*		0.6 (0.7)	11.2 (11.8)
Extant humans (n=10)	6.9 (0.3)	1.9 (0.6)	1.3 (0.2)	3.8 (0.5)	52.5 (7.6)
Extant chimpanzees (n=10)	4.6 (0.6)	3.9 (0.6)*		0.8 (0.2)	14.8 (1.9)

**The outer and inner tables are indistinct (i.e. cortical) in SK 46 and extant chimpanzees.*

Table 3: Mean frontal bone thickness (FBT, mm) of human and non-human hominin species from Copes and Kimbel[16]. Standard deviations are shown in brackets.

	n	FBT	Outer table	Inner table	Diploë
Australopithecus afarensis	2	7.5 (1.8)	1.4 (0.4)	1.3 (0.3)	6.0 (0.3)
Paranthropus boisei	7	6.0 (1.2)	–	–	–
Early *Homo*	10	6.0 (1.5)	–	–	–
European *Homo erectus*	2	8.5 (0.7)	–	–	–
African *Homo erectus*	8	9.2 (1.3)	–	–	–
Asian *Homo erectus*	29	9.2 (2.1)	–	–	–
Homo heidelbergensis	14	8.5 (1.8)	–	–	–
Homo neanderthalensis	13	6.6 (1.6)	–	–	–
Pleistocene *Homo sapiens*	17	7.5 (2.1)	–	–	–

The mean frontal bone thickness in StW 671 is thick compared to Sts 5, Sts 71, SK 46 and the extant human and chimpanzee specimens with the exception of StW 505 (Table 1, Figure 4a). When compared to the frontal bone thickness values reported by Copes and Kimbel[16], the closest matches are *Homo erectus* and *Homo heidelbergensis* (Table 3). StW 671 shares the following pattern of bone tissue proportions with Sts 71, extant humans and *Australopithecus afarensis*: the diploë is the thickest bony layer while the inner table is the thinnest (Tables 1 and 3). As previously noted[20,21], Sts 5 lacks a significant portion of its outer table. In terms of tissue proportions (Table 1, Figure 4a), the diploic bone contributes 45.1% to the cumulative cranial thickness in StW 671, while in Sts 5, Sts 71 and extant humans the diploë represents more than 50% of the total thickness and less than 20% in SK 46 and extant chimpanzees (Figure 4b).

The mean occipital bone thickness in StW 672 is thin compared to the comparative fossil specimens and extant humans (Table 2, Figure 4c). The pattern of bone tissue proportions in the comparative sample is variable but StW 672 and Sts 5 and Sts 71 have a similar pattern in that the diploë is the thickest bony layer while the outer table is the thinnest (Table 2). The proportion of diploë in StW 672 represents 41.3% of the cumulative cranial thickness and more than 50% in StW 578, Sts 5, Sts 71 and extant humans but less than 20% in SK 46 and extant chimpanzees (Figure 4d).

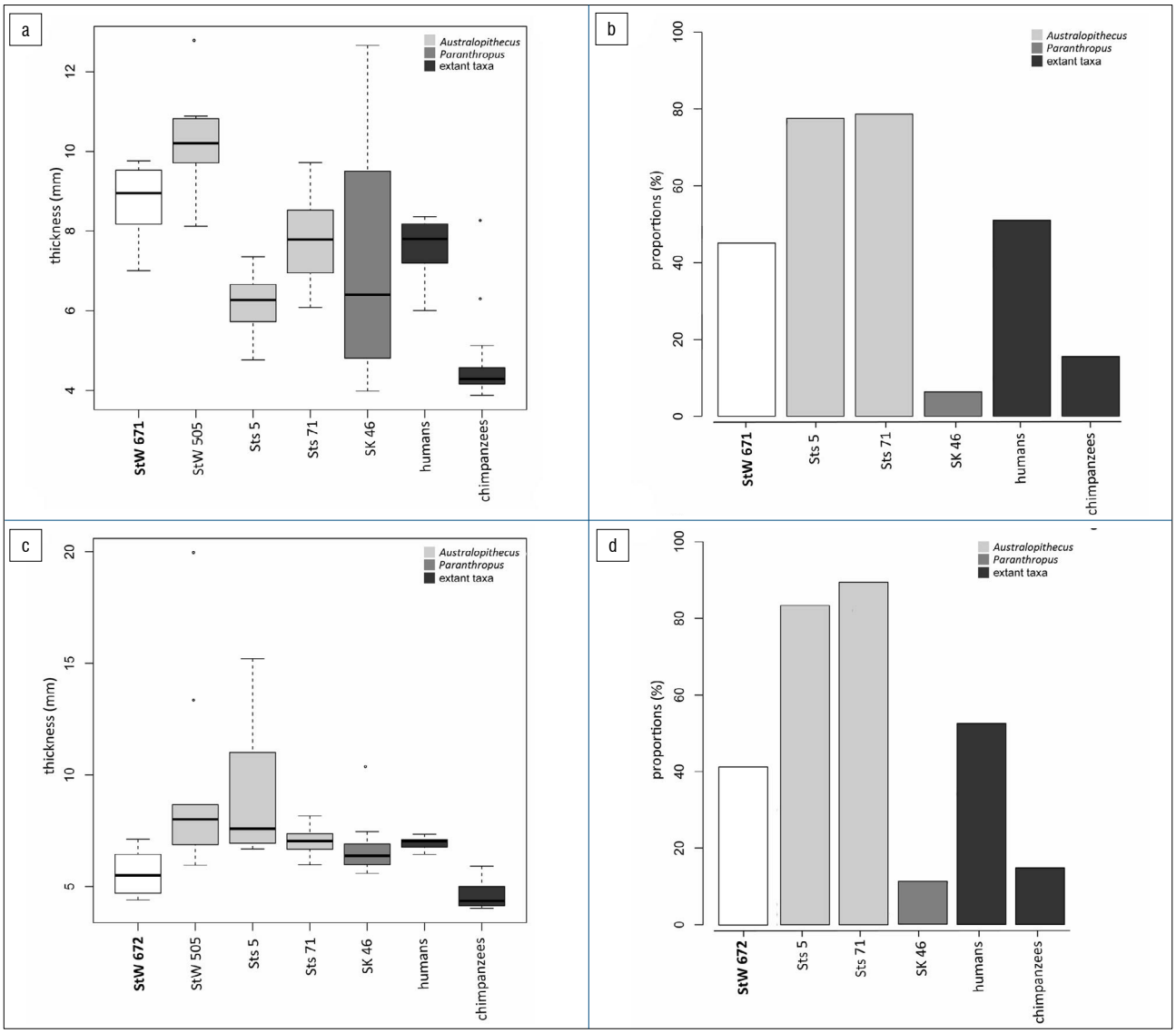


Figure 4: (a and c) Box plots of cranial thickness and (b and d) histograms of proportion of diploë in (a, b) StW 671 and (c, d) StW 672 compared to some Plio-Pleistocene hominin specimens/taxa and extant human crania.

Discussion

While the overall morphology and cranial thickness of StW 671 approximates the human condition, the proportion of diploë in both StW 671 and StW 672 is lower than that found in human and non-human hominin taxa investigated in this study with the exception of *Paranthropus*. Interestingly, despite StW 671 showing a derived human-like cranial morphology and thickness, the proportions of diploë in StW 671 and StW 672, even if comparatively closer to extant humans, do not clearly match the fossil hominins included in our comparative study. In this context, our analysis may suggest a combination of derived human-like traits and unique features. This mosaicism is compatible with the description and metric analyses of the manual proximal phalanx StW 668 and the upper right first molar StW 669 from Milner Hall and contributes to a certain degree of taxonomic ambiguity.[1] As the diploic bone acts as a protective barrier for the brain, contributes to mechanical properties of the cranium and plays a role in the cranial vascular system, our results may potentially suggest an intriguing distinct palaeobiology in the Milner Hall hominins.[16,22-25]

Further analyses documenting variation in the frontal anatomy (e.g. post-orbital constriction) and cranial thickness in early hominins, and more particularly in South African and East African early *Homo*, would be critical for the interpretations of the Milner Hall hominin palaeobiology and taxonomic attribution. Moreover, comparisons with the East African early *Homo* material preserving frontal and occipital bones (e.g. KNM-ER 1470, KNM-ER 1813) would be particularly relevant for supporting the presence of human-like traits in the Milner Hall hominins' anatomy and/or assessing the potential uniqueness of the MH fossil record. Additionally, because the position of our sections might vary across specimens, our bidimensional quantitative investigation of the cranial bone thickness and composition should be combined in future with 3D approaches[16] and/or supported by new findings of more complete cranial fragments from the Milner Hall deposits. Nevertheless, besides reporting additional specimens from STK-MH1, this study highlights the relevance of the analysis of cranial thickness and composition in taxonomical studies and the potential of the Milner Hall for contributing to our knowledge of the hominin palaeobiodiversity at Sterkfontein.

Acknowledgements

We are indebted to E. Gillisen and W. Wendelen (Tervuren), G. Krüger (Pretoria), L. Kgasi, H. Fourie, S. Potze and M. Tawane (Pretoria) and B. Zipfel (Johannesburg) for having granted access to fossil and comparative material under their care. We also thank L. Bam, F. de Beer and J. Hoffman (Pelindaba), M. Dierick (Ghent) and K. Jakata (Johannesburg) for microtomographic acquisitions. We are grateful to the Ditsong National Museum of Natural History and the University of the Witwatersrand for loaning hominin crania in their collection. For technical and/or scientific discussion/collaboration we are grateful to: M. Carmen Arriaza (Johannesburg), R. Clarke (Johannesburg), J. Dumoncel (Toulouse), K. Carlson (Los Angeles) and A. Oettlé (Pretoria). We thank the DST-NRF for sponsoring the Micro-XCT facility at Necsa, and the DST-NRF and the University of the Witwatersrand for funding the microfocus X-ray CT facility in the ESI (www.wits.ac.za/microct). The support of the AESOP+ Programme, the Claude Leon Foundation, the DST-NRF Centre of Excellence in Palaeosciences, the French Institute of South Africa and the Palaeontological Scientific Trust (PAST) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the Centre of Excellence in Palaeosciences.

Authors' contributions

Conceptualisation: A.B., T.R.P., J.L.H., D.S.; methodology: A.B., E.N.L.; data collection: A.B., D.S., T.R.P., J.L.H.; sample and data analysis: A.B., E.N.L., T.R.P., J.L.H.; validation: T.R.P., J.L.H., E.N.L.; writing: A.B., D.S., T.R.P., J.L.H., E.N.L.; project leadership: D.S.; project management: A.B.; funding acquisition: A.B., D.S.

References

1. Stratford D, Heaton JL, Pickering TR, Caruana MV, Shadrach K. First hominin fossils from Milner Hall, Sterkfontein, South Africa. J Hum Evol. 2016;91:167–173. https://doi.org/10.1016/j.jhevol.2015.12.005
2. Stratford DJ, Grab S, Pickering TR. The stratigraphy and formation history of fossil-and artefact-bearing sediments in the Milner Hall, Sterkfontein Cave, South Africa: New interpretations and implications for palaeoanthropology and archaeology. J Afr Earth Sci. 2014;96:155–167. https://doi.org/10.1016/j.jafrearsci.2014.04.002
3. Granger DE, Gibbon RJ, Kuman K, Clarke RJ, Bruxelles L, Caffee MW. New cosmogenic burial ages for Sterkfontein Member 2 *Australopithecus* and Member 5 Oldowan. Nature. 2015;522:85–88. https://doi.org/10.1038/nature14268
4. Clarke RJ. *Australopithecus* from Sterkfontein Caves, South Africa. In: Reed K, Fleagle JG, Leakey RE, editors. The paleobiology of *Australopithecus*. Dordrecht: Springer; 2013. p. 105–123. https://doi.org/10.1007/978-94-007-5919-0
5. Grine FE. The alpha taxonomy of *Australopithecus africanus*. In: Reed KE, Fleagle JG, Leakey RE, editors. The paleobiology of *Australopithecus*. Dordrecht: Springer; 2013. p. 74–104. https://doi.org/10.1007/978-94-007-5919-0
6. Grine FE, Smith HF, Heesey CP, Smith EJ. Phenetic affinities of Plio-Pleistocene *Homo* fossils from South Africa: Molar cusp proportions. In: Grine FE, Fleagle JG, Leakey RE, editors. The first humans: Origin and early evolution of the genus *Homo.* Dordrecht: Springer; 2009. p. 49–62. https://doi.org/10.1007/978-1-4020-9980-9
7. Curnoe D. A review of early *Homo* in southern Africa focusing on cranial, mandibular and dental remains, with the description of a new species (*Homo gautengensis* sp nov.). Homo. 2010;61:151–177. https://doi.org/10.1016/j.jchb.2010.04.002
8. Clarke RJ. *Homo habilis*: The inside story. In: Sahnouni M, Semaw S, Garaizar JR, editors. Proceedings of the II Meeting of African Prehistory. Burgos: Consorcio CENIEH; 2017. p. 25–51.
9. Clarke RJ. The cranium of the Swartkrans hominid, SK 847 and its relevance to human origins [PhD thesis]. Johannesburg: University of the Witwatersrand; 1977.
10. Broom R. Discovery of a new skull of the South African ape-man, *Plesianthropus.* Nature. 1947;159:672. https://doi.org/10.1038/159672a0
11. Broom R, Robinson JT. Further evidence of the structure of the Sterkfontein ape-man *Plesianthropus*. Transv Mus Mem. 1950;4:7–83.
12. Broom R, Robinson JT. Swartkrans ape-men. *Paranthropus crassidens*. Pretoria: Transvaal Museum; 1952.
13. Broom R, Robinson JT, Schepers GWH. Sterkfontein ape-man, *Plesianthropus*. Transv Mus Mem. 1950;4:1–117.
14. Lockwood CA, Tobias PV. A large male hominid cranium from Sterkfontein, South Africa, and the status of *Australopithecus africanus*. J Hum Evol. 1999;36:637–685. https://doi.org/10.1006/jhev.1999.0299
15. Berger LR, Hawks J, De Ruiter DJ, Churchill SE, Schmid P, Delezene LK, et al. *Homo naledi*, a new species of the genus *Homo* from the Dinaledi Chamber, South Africa. eLife. 2015; 4, e09560, 35 pages. https://doi.org/10.7554/eLife.09560
16. Copes LE, Kimbel WH. Cranial vault thickness in primates: *Homo erectus* does not have uniquely thick vault bones. J Hum Evol. 2016;90:120–134. https://doi.org/10.1016/j.jhevol.2015.08.008
17. L'Abbé EN, Loots M, Meiring JH. The Pretoria Bone Collection: A modern South African skeletal sample. Homo. 2005;56:197–205. https://doi.org/10.1016/j.jchb.2004.10.004
18. Beaudet A, Carlson KJ, Clarke RJ, De Beer F, Dhaene J, Heaton JL, et al. Cranial vault thickness variation and inner structural organization in the StW 578 hominin cranium from Jacovec Cavern, South Africa. J Human Evol. 2018;121:201–220. https://doi.org/10.1016/j.jhevol.2018.04.004
19. Aiello L, Dean C. An Introduction to human evolutionary anatomy. New York: Academic Press; 1990. https://doi.org/10.1016/C2009-0-02515-X
20. Wolpoff MH. Sagittal cresting in the South African australopithecines. Am J Phys Anthropol. 1974;40:367–408. https://doi.org/10.1002/ajpa.1330400312

21. Thackeray JF. Cranial bone of 'Mrs Ples' (Sts 5): Fragments adhering to matrix. S Afr J Sci. 1997;93:169–170.

22. McElhaney JH, Fogle JL, Melvin JW, Haynes RR, Roberts VL, Alem NM. Mechanical properties of cranial bone. J Biomech. 1970;5:495–496. https://doi.org/10.1016/0021-9290(70)90059-X

23. Hershkovitz I, Greenwald C, Rotschild B, Latimer B, Dutour O, Jellema LM, et al. The elusive diploic veins: Anthropological and anatomical perspective. Am J Phys Anthropol. 1999;108:345–358. https://doi.org/10.1002/(SICI)1096-8644(199903)108:3<345::AID-AJPA9>3.0.CO;2-S

24. Anzelmo M, Ventrice F, Barbeito-Andres J, Pucciarelli HM, Sardi ML. Ontogenetic changes in cranial vault thickness in a modern sample of *Homo sapiens*. Am J Hum Biol. 2015;27:475–485. https://doi.org/10.1002/ajhb.22673

25. Rangel de Lázaro G, De la Cuétara JM, Píšová H, Lorenzo C, Bruner E. Diploic vessels and computed tomography: Segmentation and comparison in modern humans and fossil hominids. Am J Phys Anthropol. 2016;159:313–324. https://doi.org/10.1002/ajpa.22878